Dynamique d'Expression d'un Réseau de Régulation Génétique : la Décision Lyse/Lysogénie chez le Bactériophage Lambda

Stéphane Ghozzi

École Normale Supérieure
Laboratoire de Physique Statistique

Thèse de Doctorat de
l'Université Pierre et Marie Curie

Spécialité Physique

présentée par

Stéphane Ghozzi

pour obtenir le grade de

Docteur de l'Université Pierre et Marie Curie

Dynamique d'Expression d'un Réseau de Régulation Génétique : la Décision Lyse/Lysogénie chez le Bactériophage Lambda

Soutenance prévue le 16 décembre 2009 devant le jury composé de :

Axel Buguin	Examinateur
Gilles Charvin	Examinateur
Didier Chatenay	Directeur de thèse
Bahram Houchmandzadeh	Rapporteur
Mathias Springer	Examinateur
Denis Thieffry	Rapporteur

Résumé

Lors d'une infection par le virus Lambda, une bactérie peut suivre l'une ou l'autre de deux voies très différentes : la lyse, où le virus se multiplie, tue l'hôte et est relâché dans l'environnement, ou la lysogénie, où l'ADN viral est intégré dans le chromosome de l'hôte et celui-ci devient immunisé à une surinfection. La « décision » entre la lyse et la lysogénie est prise aléatoirement, mais les probabilités de suivre chacune des deux voies dépend du nombre de virus infectant la cellule en même temps et de son état physiologique : un faible nombre de virus par bactérie, un milieu riche ou des cellules en phase de croissance exponentielle favorisent la lyse, alors que les conditions opposées favorisent la lysogénie. Cela est rendu possible par un ensemble de cinq gènes, sous quatre promoteurs, interagissant les uns avec les autres et avec des protéines de l'hôte.

Des paires de gènes codant pour des protéines fluorescentes ont été insérés dans le génome de Lambda. En mesurant les niveaux de fluorescence de bactéries individuelles, nous pouvons déterminer, au cours de la décision, les taux d'expression de gènes de ce réseau et leurs corrélations deux à deux. En variant systématiquement les conditions de croissance, nous pouvons ainsi obtenir une description riche de la dynamique, notamment du rôle du bruit stochastique et de son contrôle, de ce réseau de régulation naturel.

Enfin, nous avons complété un programme informatique développé au laboratoire de façon à pouvoir comparer le réseau de Lambda à des réseaux générés sur ordinateur et ayant le même comportement. Cela aidera à mieux comprendre la structure particulière de ce réseau.

Abstract

Upon infection by the virus Lambda, an *E. Coli* bacterium can follow either one of two drastically different paths : lysis, where the virus multiplies, kills the host and is released in the environment, or lysogeny, where the viral DNA is integrated in the bacterial chromosome and the host becomes immunized to surinfection. The "decision" between lysis and lysogeny is taken randomly, but the probabilities of following each path depends on the number of viruses infecting the cell at the same time and the physiological state of the bacterium : a small number of viruses per bacterium, a rich medium or cells in exponential growth phase favor the lysis, while the opposite conditions favor the lysogeny. This is made possible by an elegant, compact set of five co-regulated genes, under four promoters, interacting with a few proteins of the host. Stochastic fluctuations in Lambda and host protein numbers, and their control, are thought to play a major role in the decision process.

We inserted a pair of genes coding for fluorescent proteins in this Lambda regulation region, that are co-transcribed with a pair of regulating or regulated genes, and transformed *E. Coli* bacteria with this construction cloned in a plasmid. Time-lapse fluorescent microscopy allows us to directly measure the expression dynamics and correlations of the chosen genes. Three constructions, with different pairs of tagged genes, have been made. These measurements are repeated in different conditions, by changing the growth medium and phase of the bacteria, and using plasmids of different copy numbers.

We also worked on a computer program first written by P. François and V. Hakim at LPS, that simulates directed evolution of genetic regulatory networks and allows one to design networks having a defined behavior. We have implemented detailed molecular processes, as those encountered during the lysis/lysogeny decision, and will generate networks that are able to fulfill the same functions as Lambda's (simply put, a biased bistable switch). By comparing their structure and properties with the experimental results, we hope to refine our understanding of this natural genetic regulatory network.

Table des matières

Introduction

Au sein de chaque cellule, l'ADN porte l'information nécessaire à son intégrité et à sa reproduction, ou à celle de l'organisme dont elle fait partie. Des portions d'ADN, les gènes, sont « exprimés », c'est-à-dire qu'ils sont transcrits en ARN, eux-mêmes traduits en protéines : à chaque gène est ainsi associée une protéine remplissant une ou plusieurs fonctions déterminées, souvent de catalyse. Certaines de ces protéines interagissent avec l'ADN, des ARN ou d'autres protéines de façon à en modifier la concentration. On peut se représenter ces intéractions comme un réseau de gènes, où l'expression de l'un modifie l'expression d'un ou plusieurs autres.

Ces réseaux de régulation permettent une réponse coordonnée de l'expression des gènes adaptée à un environnement ou au programme de développement d'un organisme. Ce contrôle rappelle par certains aspects les circuits utilisés en électronique pour traiter des signaux. En particulier, par exemple, des boucles de rétroaction positives permettent un fonctionnement bimodal de la cellule adapté à un environnement : deux ensembles de gènes distincts seront exprimés suivant que la concentration d'une molécule (signal) sera supérieure ou inférieure à un seuil.

Ils assurent aussi les comportements dynamiques seuls capables de maintenir la vie : croissance, reproduction du génome, division cellulaire, qui doivent être réguliers et coordonnés. Les cycles circadiens sont un autre exemple de comportement dynamique possible uniquement par un ensemble de réactions chimiques couplées non-linéaires.

Un faible nombre de molécules interviennent dans l'expression d'un gène : celle-ci est sujette, du fait de l'agitation thermique, à d'importantes fluctuations [1, 2]. Ce bruit peut *a priori* nuire au fonctionnement cellulaire, en ruinant la stabilité d'un schéma d'expression, sa cohérence temporelle ou la coordination de l'expression de plusieurs gènes [3]. À ce bruit propre aux processus intracellulaires s'ajoutent les fluctuations de l'environnement, qui peuvent être particulièrement importantes pour des bactéries par exemple.

S'il est certain que les réseaux de régulation génétiques doivent être adaptés à ces fluctuations, les principes de leur contrôle sont encore mal compris [4, 5, 6]. Il a été proposé qu'un trait caractéristique des réseaux biochimiques soit leur robustesse face à des variations de leurs constantes cinétiques [7, 8]. Récemment, la manière dont le bruit se propage

dans une cascade de régulation génétique [9] et les corrélations d'expression de gènes de petits réseaux de régulation [10] ont pu être mesurées.

Mais aucune description expérimentale n'existe, pour un réseau complexe, du traitement de telles fluctuations. Le but de cette thèse a été de fournir une telle description.

L'utilisation de gènes codant pour des protéines fluorescentes permet de suivre en temps réel l'expression de gènes d'intérêt. Des mesures de fluorescence sur bactéries uniques permettent non-seulement de distinguer plusieurs modes d'expression au sein d'une population de cellules génétiquement identiques et placées dans le même environnement, mais aussi d'accéder aux fluctuations d'expression génétique.

J'ai choisi d'étudier un réseau de régulation bien connu, celui responsable de la décision lyse/lysogénie chez le bactériophage Lambda.

Le chapitre suivant expose les traits essentiels du fonctionnement de ce réseau et l'intérêt qu'il y a à l'étudier dans ce contexte. Le détail des expériences menées et des résultats obtenus sont donnés au chapitre 2 (les matériels et méthodes étant décrits en annexe A). Le chapitre 3 présente l'algorithme d'évolution sur lequel j'ai travaillé et les développements que j'y ai apportés.

J'ai inclus dans ce manuscrit, en annexe B, un travail sur le nombre de copies de plasmides.

Chapitre 1

Lambda : un modèle de réseau de régulation génétique

Lambda est un virus de la bactérie (un « bactériophage ») *Escherichia coli*. Il est, depuis sa découverte dans les années 50, un modèle d'étude du fonctionnement de l'expression génétique et de son contrôle. Je présente dans la suite sommairement le fonctionnement du réseau de régulation génétique qui nous intéressera (section 1.1), puis les questions qu'il peut encore nous aider à poser et la manière dont j'ai choisi de l'étudier (section 1.2). Pour une présentation plus complète, on pourra consulter en particulier [11], [12] et [13], qui contient un chapitre clair et concis sur Lambda.

Le tableau 1.1 présente un glossaire de termes couramment utilisés.

1.1 Présentation du réseau

Je présente dans la suite les cycles de vie (voies de développement) de Lambda, la manière dont il peut « choisir » entre lyse et lysogénie et les conditions influençant ce choix, et les mécanismes de maintien et de sortie de lysogénie. Trois échelles seront considérées au cours de l'exposé : cellulaire (paragraphe 1.1.1), du réseau de gènes (paragraphe 1.1.2) puis d'un ou deux gènes en particulier (paragraphes « La compétition Cro/CI en O_R » et « Le rôle pivot de CII », puis 1.1.3). De nombreux détails de mode d'action des protéines régulatrices seront passés sous silence pour ne pas alourdir l'exposé.

1.1.1 Les voies de développement

Lors de l'infection d'une cellule par Lambda, le virion s'accroche à la surface cellulaire et y injecte son ADN. L'ADN se circularise, et certains de ses gènes commencent à être exprimés par la « machinerie » (polymérases, ribosomes, etc.) de l'hôte. L'une ou l'autre de deux voies peut alors être suivie : lors de la *lyse*, l'ADN viral est répliqué, les protéines formant les capsides produites, les virions assemblés et enfin les parois de la cellule hôte

virion ou particule virale	enveloppe (la capside) et l'ADN qu'elle contient ; c'est la forme du virus dans l'environnement.
lyse	dans un sens restreint, dégradation de la paroi cellulaire ; dans un sens large, mode d'infection où le virus utilise la cellule hôte pour se multiplier ; la cellule est alors détruite (paroi « lysée »).
lysogénie	mode d'infection où l'ADN viral est inséré dans le chromosome de l'hôte, immunisant celui-ci à une surinfection par Lambda, et est répliqué avec lui.
induction (de la lyse)	levée de la lysogénie, excision de l'ADN viral du chromosome, développement suivant la voie lytique.
prophage	ADN viral inséré dans le chromosome ; qualifié de défectif si une lyse ne peut plus être induite, c'est-à-dire que l'ADN viral ne peut plus être excisé du chromosme.
lysogène	bactérie contenant un prophage (sous-entendu non défectif).
phage tempéré	phage capable de suivre l'une ou l'autre des deux voies, lyse et lysogénie.
répresseur	protéine maintenant la lysogénie et empêchant la surinfection : CI chez Lambda.
région d'immunité (du génome de Lambda)	portion compacte du génome de Lambda responsable à la fois de la décision lyse/lysogénie et de l'immunité à la surinfection conférée au lysogène (voir détail sur la figure 1.2 et sa légende).

Tab. 1.1 – Glossaire relatif à Lambda de termes couramment utilisés.

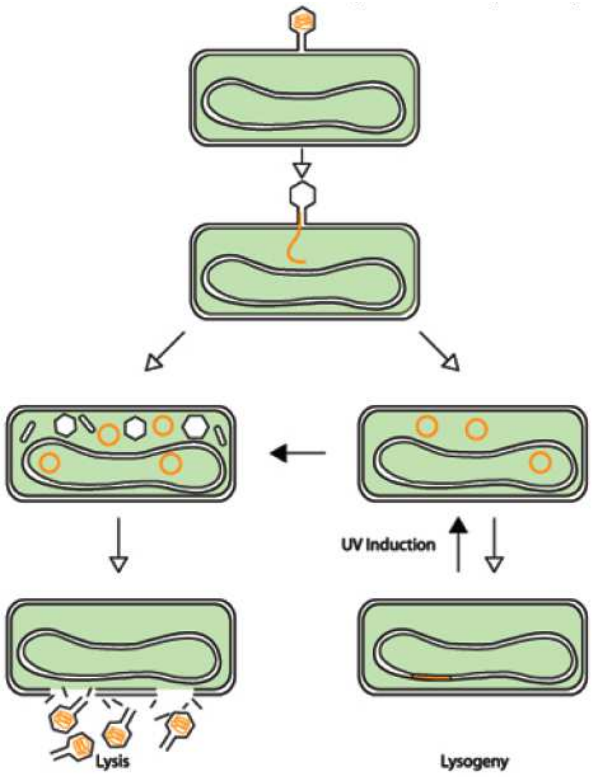


Fig. 1.1 – Schéma des voies de développement pouvant être suivies à l'infection d'une bactérie *E. coli* par le phage Lambda. En orange, l'ADN viral.

détruites et les virions relâchés dans l'environnement ; lors de la *lysogénie*, l'ADN viral s'intègre dans le chromosome de l'hôte au site *attB*, et l'hôte est rendu immunisé à une surinfection par Lambda. Enfin, la lyse peut être induite chez un lysogène : l'ADN viral est alors excisé du chromosome, se circularise et suit le développement lytique décrit précédemment. La figure 1.1 présente un schéma de ces processus.

Chacune de ces étapes est médiée par des protéines encodées dans des gènes viraux, éventuellement en coopération avec des protéines de l'hôte. La figure 1.2 présente la carte génétique de Lambda. La structure modulaire de ce génome est frappante : les gènes participant d'une même fonction sont voisins, et forment des groupes connexes sur le génome. Les opérons (ensemble de gènes sous un même promoteur) du génome reproduisent à peu près cette organisation, mais peuvent être particulièrement intriqués, en particulier dans la région des bases 33000 à 39000, et autour du gène *Q* (autour de la base 44000).

L'ADN viral est linéaire d'extrémités indiquées sur la figure 1.2 (ouvert au site COS) dans la capside, circulaire dans le cytoplasme, et linéaire ouvert en *attP* (près de la base 28000) quand il est inséré dans le chromosome de l'hôte.

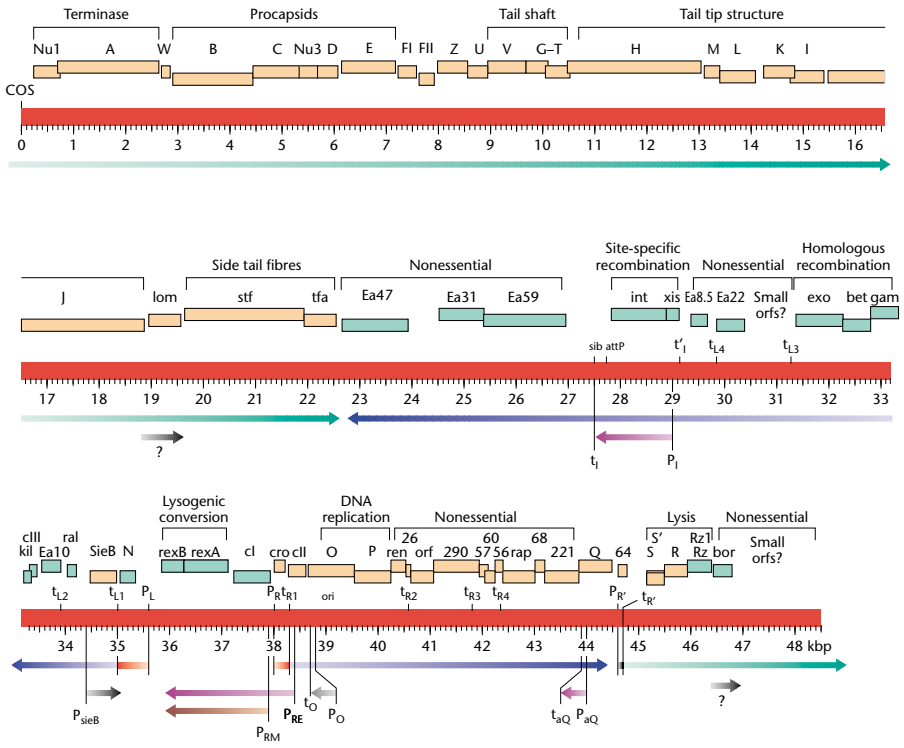


Fig. 1.2 – Carte génétique du bactériophage Lambda. La « région d'immunité », responsable de la décision lyse/lysogénie, se trouve en bas à gauche de la carte, du gène *cIII* au promoteur P_O (bases 33299 à 38599). Les flèches indiquent les transcrits pouvant être produits. En orange, les gènes d'orientation directe (vers la droite sur cette carte), en vert ceux d'orientation inverse. *Remarque* : P_O sera nommé P_{OOP} dans la suite, suivant la convention la plus répandue. Figure tirée de [14].

1.1.2 La décision lyse/lysogénie

Interactions génétiques

Ce qui suit est une présentation des processus biochimiques qui conduisent à la lyse puis de ceux qui conduisent à la lysogénie, respectivement représentés schématiquement sur les figures 1.3 et 1.4. J'ai choisi un ordre d'exposé qui rende claire leur *logique* ; l'apparence d'ordre *temporel* pourra être trompeuse. Je reviendrai plus en détail dans les paragraphes suivants sur les modes d'action de trois protéines virales clés : Cro, CI et CII.

Les seuls promoteurs constitutifs de Lambda sont P_R et P_L, ils sont donc les seuls actifs à l'infection [1]. Ils conduisent à la transcription des gènes N et *cro*, et à un taux plus faible de *cII* : les terminaisons de transcription t_{L1} et t_{R2} empêchent les gènes suivant d'être transcrits et t_{R1} n'a qu'une efficacité de 50% (figure 1.3 A) [2]. Cro dimérise et ce dimère peut se fixer au site O_{R3}, qui recouvre le promoteur P_{RM}... qui n'est cependant pas actif (plus de détails sont donnés au paragraphe suivant « La compétition Cro/CI » et sur la figure 1.5). N, agissant avec des protéines de l'hôte, lève les terminaisons de transcription, permettant l'expression forte de *cII* et la transcription des autres gènes en aval des promoteurs P_R et P_L (figure 1.3 B), en particulier les gènes O et P, responsables de la réplication de l'ADN viral, puis le gène Q, antiterminateur de $t_{R'}$ derrière laquelle se trouvent les gènes de la lyse et de la capside. Ainsi, en l'absence d'autres processus, la lyse sera suivie. Nous verrons ci-dessous que CII peut inhiber la lyse, mais, bien qu'il soit maintenant fortement exprimé, des protéines de l'hôte peuvent réprimer son action : HflB [3] et la RNaseIII. Elles favorisent ici toutes deux la lyse, la première en dégradant CII, la seconde en dégradant le double-brin d'ARN formé sur le transcrit de *cII* sur lequel peut venir s'hybrider un petit ARN produit sous P_{OOP}. La RNase III dégrade aussi une boucle formée par le transcrit produit sous P_L après passage de t_{L1}, ce qui permet de maintenir la traduction de N, et donc l'expression des gènes lytiques (figure 1.3 C). À haute concentration, Cro peut inhiber de deux façons l'activité de CII : en se fixant sur O_{R1}, il réprime l'activité de P_R et donc la transcription de *cII*, en se fixant sur O_{L1}, il réprime l'activité de P_L, et donc la production de CIII qui stabilise CII (figure 1.3 D).

Considérons maintenant *cIII* ainsi que l'action possible de CII, qui a commencé à s'accumuler. Tous deux favorisent la lysogénie : voyons leur action dans le cas où la lysogénie est suivie (figure 1.4 A). CIII protège CII de la dégradation par HflB ; de plus, la RNa-

1. C'est aussi le cas de $P_{R'}$, mais les transcriptions qui y sont initiées sont immédiatement terminées en $t_{R'}$.

2. On remarquera que les gènes O et P, responsables de la reproduction de l'ADN viral, sont aussi transcrits ; ce début de réplication joue-t-il un rôle dans la décision ? Les auteurs de [15], constatant que la lysogénie est majoritairement suivie dès que le réseau est présent en plus de deux copies, suggèrent que non.

3. Aussi appelée FtsH, ayant été découverte séparément dans des contextes différents.

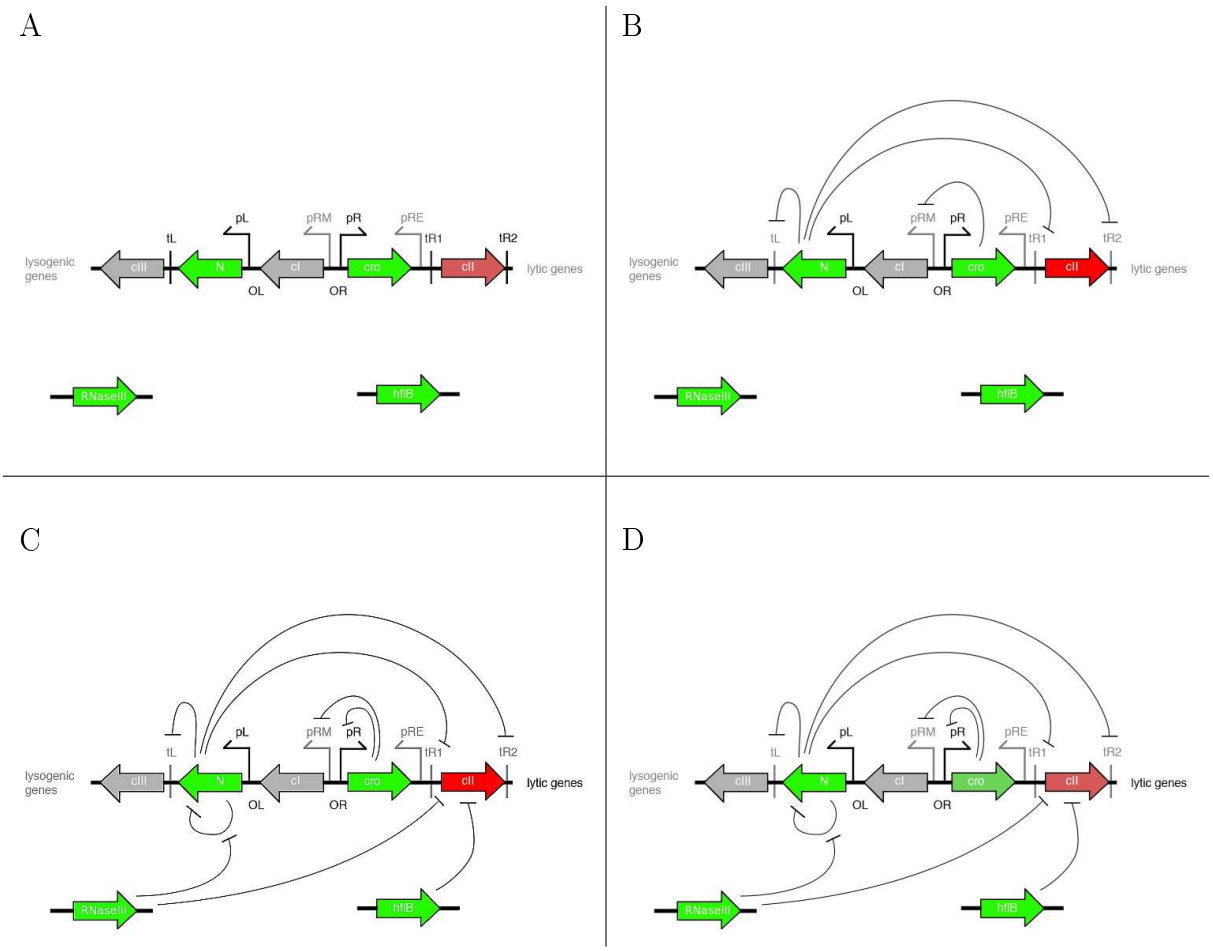


Fig. 1.3 – Schémas des cascades de régulations conduisant à la lyse. Seuls les gènes, promoteurs et terminaisons y participant sont indiqués (à l'exception de P_{OOP}). En couleur, les gènes exprimés (verts s'ils favorisent la lyse, rouges s'ils favorisent la lysogénie). Les flèches pointues indiquent une action qui favorise l'activité du gène vers lequel elle pointe, les flèche barrées une inhibition ; les modes d'action qu'elles recouvrent peuvent être très variés (voir le texte principal). Pour alléger les dessins, le détail des sites O_R et O_L n'est pas reproduit et une terminaison t_L a été indiquée au lieu des deux terminaisons t_{L1} et t_{L2}. *Remarques* : 1. par souci de lisibilité, « RNaseIII » est indiqué sur les dessins, bien que le gène dont elle est le produit s'appelle *rnc* ; 2. bien que pouvant être exprimé, le gène *cIII* n'a pas été considéré ici, son produit favorisant la lysogénie (voir la figure 1.4 et le texte principal).

seIII aide au repliement de CIII (figure 1.4 B). Or CII a une action décisive : en activant le promoteur P_{aQ} il inhibe l'expression de Q et donc des gènes lytiques ; en activant le promoteur P_I, il permet la production de Int qui réalise l'intégration dans le chromosome ; enfin, CII active le promoteur P_{RE} (figure 1.4 C). Or la transcription sous P_{RE} peut gêner la transcription de *cro*, des ARN-polymérases avançant en sens contraire le long de l'ADN (quoique sur des brins différents), et surtout permet l'expression de *cI* (figure 1.4 D). CI, sous forme de dimère, réprime très efficacement les promoteurs P_R et P_L, les protéines CII encore présentes finissant d'activer l'intégration dans le chromosome et d'inhiber l'expression des gènes lytiques (figure 1.4 E). Enfin, CI bloque l'expression de tous les autres gènes viraux (à part *rexA*, *rexB* et *sieB*, impliqués dans l'immunité à la surinfection par d'autres bactériophages), et maintient sa propre production sous P_{RM} : c'est la lysogénie (figure 1.4 F). La répression de P_R et P_L par CI assure aussi l'immunité du lysogène à une surinfection par Lambda.

Si j'ai tenté de *raconter* la manière dont l'infection peut aboutir à l'un ou l'autre mode, il faut se rappeler que ces processus peuvent *a priori* tous se produire en même temps (si ce n'est au départ un temps nécessaire à l'accumulation de N). On constate cependant que leur agencement temporel doit jouer un rôle important.

Cette esquisse permet de comprendre comment ce réseau de régulation génétique peut conduire à l'une ou l'autre voie de développement, et ce de façon exclusive[4]. Deux éléments du réseau semblent particulièrement aptes à produire un tel comportement : la compétition entre Cro et CI pour se fixer aux sites O_R et l'action de CII qui à la fois inhibe la lyse et active la lysogénie.

La compétition Cro/CI en O_R

Longtemps présentée comme le cœur du réseau, la portion *cI*-P_{RM}-P_R-*cro* semble constituer un commutateur génétique : deux gènes se répriment mutuellement, leur dimérisation et la compétition pour s'accrocher à leur opérateur augmentant la stabilité de chacun des deux états stationnaires possibles. La figure 1.5 montre le détail du site O_R et des promoteurs P_{RM} et P_R ainsi que le schéma de la régulation mutuelle de *cro* et *cI*.

Les dimères de Cro ont une affinité à l'ADN décroissante de O_{R3} à O_{R1}, et s'y fixent coopérativement [17] ; en O_{R3}, Cro bloque la transcription sous P_{RM} (à forte concentration, quand la probabilité d'occupation de O_{R1} n'est plus négligeable, Cro inhibe la transcription sous P_R). Les dimères de CI ont au contraire une affinité décroissante de O_{R1} à O_{R3}, et se fixent coopérativement sur les deux premiers : un dimère de CI en O_{R1} favorise fortement l'occupation de O_{R2} par CI_2. De plus, les dimères de CI en O_{R1} répriment la transcription sous P_R et en O_{R2} activent celle sous P_{RM}. La coopérativité de fixation des dimères

4. Il a été récemment suggéré qu'une voie « mixte » puisse être suivie [16].

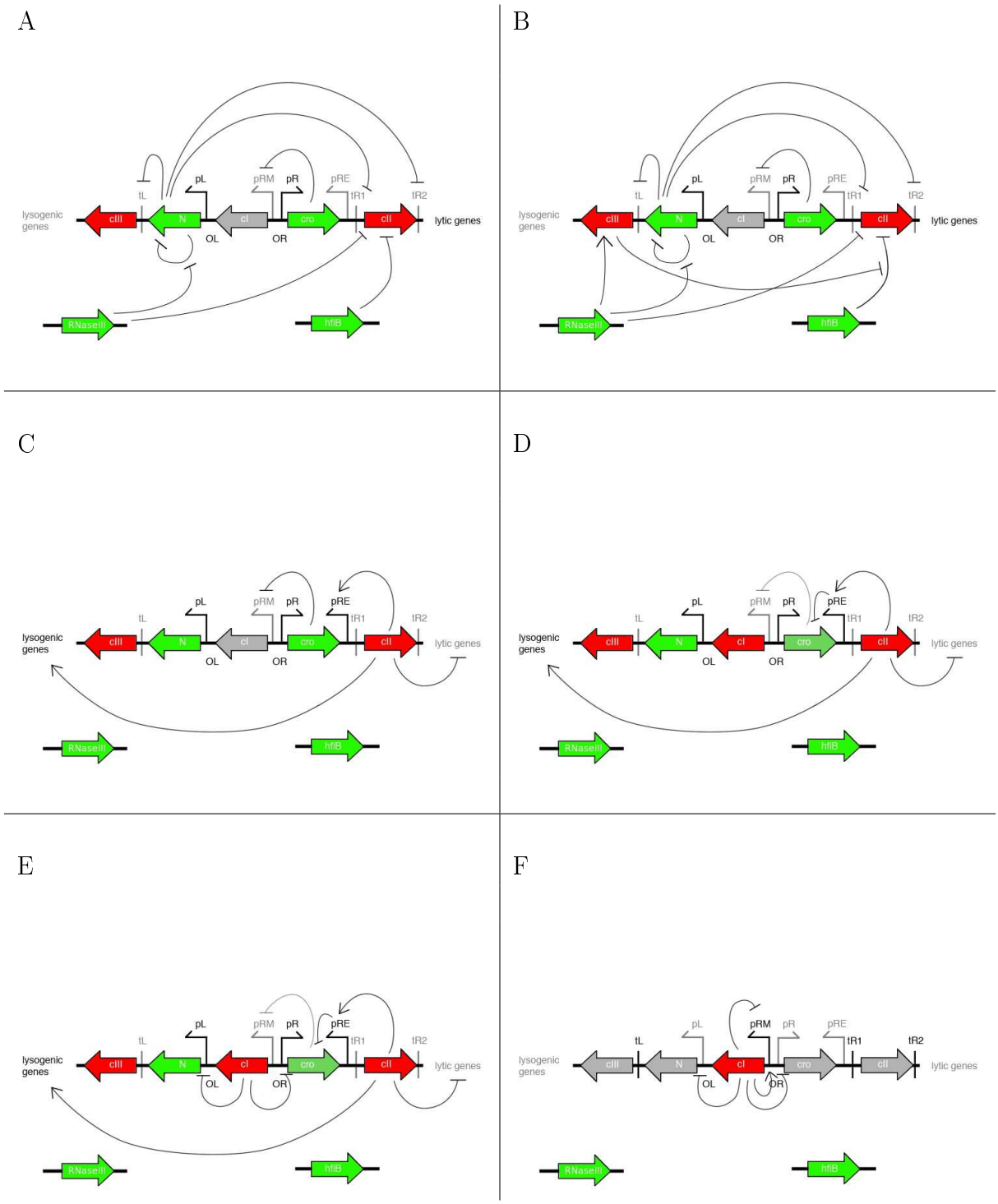
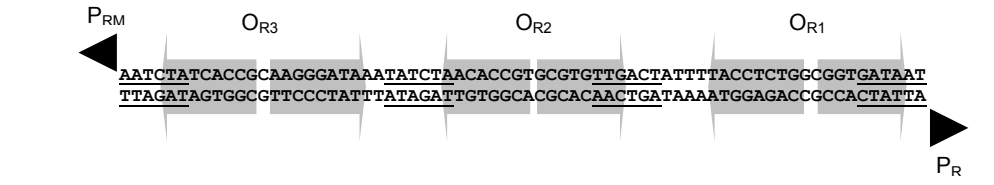


Fig. 1.4 – Schéma des cascades de régulations conduisant à la lysogénie. À partir de la figure (C), pour alléger les dessins, les actions de N, la RNaseIII et HflB ne sont plus indiquées (la production de ces deux dernières protéines dépend des conditions de croissance, voir le paragraphe « Conditions influençant la décision » de la section 1.1.2).

A



B

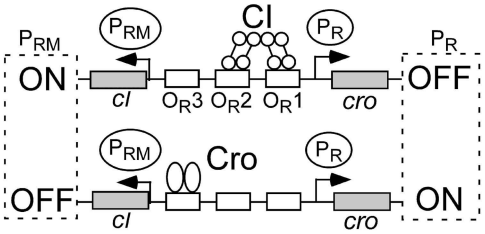


Fig. 1.5 – (A) Organisation de l'opérateur O_R de Lambda : les flèches grises représentent les sous-opérateurs impliqués dans la fixation de CI ; les séquences des boîtes -10 et -35 de P_R et P_{RM} sont soulignées ; figure tirée de [19]. (B) Les deux modes d'expression de la région *cI*-P_{RM}-P_R-*cro* ; figure tirée de [20].

augmente encore la (bi)stabilité de ce système.

Si *cro* est nécessaire à ce que la voie lytique puisse être suivie, la conception de son rôle comme essentiellement antagoniste de *cI* a récemment été remise en cause, notamment au vu de son rôle dans l'induction de la lyse (voir le paragraphe « Induction de la lyse » dans la section 1.1.3, [11, 18] et les références incluses).

Le rôle pivot de CII

Une autre approche, qui est celle que j'ai suivie dans la présentation du fonctionnement du réseau, consiste à remarquer que CII permet d'inverser une direction initiale spontanée vers la lyse, considérée comme mode « par défaut » [11], que s'il s'accumule suffisamment la lysogénie sera suivie. Il apparaît comme un nœud de régulation. Je n'ai fait que l'esquisser plus haut, la figure 1.6 montre plus en détail la place occupée par CII au sein du réseau.

La diversité des types de régulations impliquées dans le fonctionnement de CII est étonnante. *cII* est sous une terminaison de transcription t_{R1} imparfaite, qui nécessite la pr{esence du facteur d'hôte Rho pour fonctionner et est levée par l'action de N et de facteurs de l'hôte (facteurs Nus) ; le promoteur P_R sous lequel il se trouve est régulé par les facteurs de transcription Cro et CI ; sous le promoteur P_{OOP} est produit un ARN anti-sens de la fin du transcrit portant *cII*, qui s'y hybride et conduit à sa dégradation par l'intermédiaire de la RNaseIII ; CII est dégradé par HflB, mais protégée de cette dégradation par CIII ; un

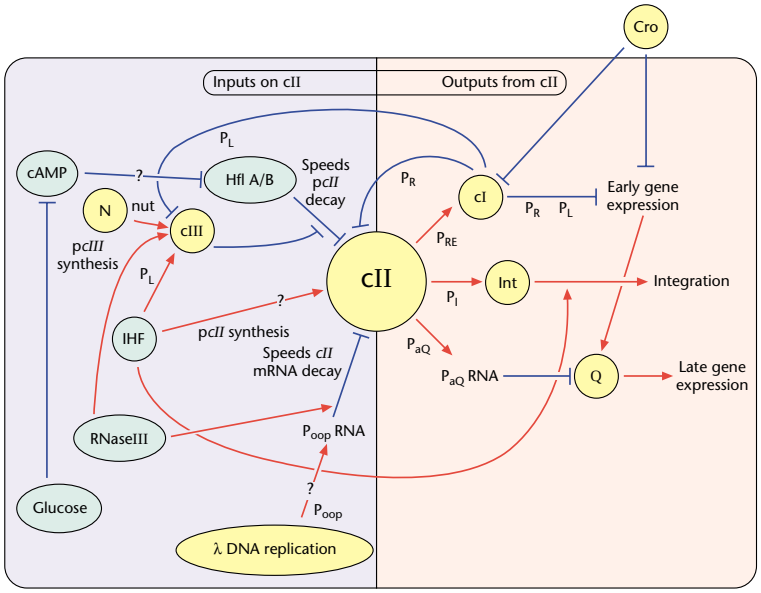


Fig. 1.6 – Schéma des interactions impliquant CII. En bleu, le facteurs d'hôte ; en jaune, les protéines virales. Figure tirée de [14].

facteur d'hôte, IHF [5], stimule sa traduction par un mécanisme inconnu. Tous ces processus modulent l'activité de CII, qui est un facteur de transcription activant trois promoteurs : P_{RE}, P_{aQ} et P_I.

Si la modulation de l'activité de CII apparaît à la fois riche et subtile, résultante de nombreux « signaux », son action est très cohérente : établir la répression (production de CI), inhiber les gènes lytiques *via* la déstabilisation du transcrit de Q, déclencher l'intégration dans le chromosome.

Conditions influençant la décision

Le tableau 1.2 présente les principales conditions ayant une influence sur le résultat de l'infection et le sens dans lequel elles influent. Elles n'ont pas chacune le même « poids » : ainsi par exemple, quelles que soient les autres conditions, une multiplicité d'infection (MOI) [6] élevée conduit toujours à une lysogénie. D'après [15], cela est vrai dès que le nombre de phages infectant une bactérie simultanément est supérieur à 2.

Dans les conditions de MOI faible, de bactéries en croissance exponentielle dans du milieu riche, la lyse est suivie dans près de 99% des infections. On peut trouver des conditions dans lesquels la lysogénie est suivie dans environ la moitié des cas : par exemple, suivant

5. Par ailleurs critique pour l'intégration de l'ADN viral dans le chromosome ; IHF : *Integration Host Factor*.

6. La multiplicité d'infection est le nombre moyen de phages infectant une bactérie en même temps, admis égal au rapport du nombre de phages sur le nombre de bactéries dans une culture s'il est supérieur à 1, et 1 sinon ; en anglais, *multiplicity of infection*, abrégé en « MOI ».

Un milieu	riche	pauvre
Une multiplicité d'infection (MOI)	faible	élevée
Une phase de croissance de l'hôte	exponentielle	stationnaire
	favorise la **lyse**.	favorise la **lysogénie**.

Tab. 1.2 – Principales conditions ayant une influence sur le résultat de l'infection et sens dans lequel elles influent (à comprendre « toutes choses égales par ailleurs »).

[16], en infectant à MOI de 1 des bactéries qui ont poussé en phase stationnaire dans du milieu riche.

On constate que l'influence de ces conditions fait sens dans une perspective écologique. Plus il y a de ressources disponibles pour les bactéries, mieux elles croissent et moins il y a de virions dans l'environnment, moins il y a de chance que toutes les bactéries soient tuées par lyse et que Lambda ne puisse plus se reproduire : dans ces conditions, la lyse, mode de reproduction efficace, sera préférée. Inversement, la lysogénie permet au virus de sortir, avec son hôte, d'une niche peu favorable et assure que quelles que soient les conditions, au moins une bactérie survivra. Comme mentionné plus haut et décrit en détail plus loin, il existe enfin un mécanisme de sortie de lysogénie, déclenché quand le chromosome de l'hôte est abîmé, ce qui permet à terme un reprise de la reproduction lytique [7].

Notons ici une découverte récente : les bactéries persistantes [8] ont une fréquence de lysogénisation plus forte que les non-persitantes [21].

Au niveau moléculaire, on peut comprendre en partie l'influence du milieu par HflB et la RNaseIII. Les taux de production de HflB et de la RNaseIII suivent le taux de croissance, si bien qu'elles sont toutes deux exprimées en milieu riche et peu ou pas en milieu pauvre. L'action de la RNaseIII sur l'expression de N explique que son taux d'expression suive la même tendance. D'autres facteurs d'hôtes interviennent : le facteur Rho pour les terminaisons, les facteurs Nus pour les antiterminaisons, le « facteur d'intégration » IHF. Tous peuvent jouer le rôle de signaux de contextes cellulaire ou environemental.

Les auteurs de [16] suggèrent que le volume même de la bactérie à l'infection biaise la décision en changeant, à taux de production fixés, les concentrations des protéines virales produites initialement.

7. On peut aussi voir l'induction comme une tentative « désespérée » du virus de « s'en sortir », une dégradation de l'ADN de l'hôte pouvant être le signe de sa disparition prochaine.

8. Il s'agit d'un état physiologique dans lequel une bactérie peut resister à certains antibiotiques ; elles croissent plus lentement que des bactéries « normales ».

La manière dont le réseau « sent » son nombre de copie n'est pas comprise, quoiqu'elle est qualitativement bien reproduite par des simulations stochastiques [22]. Il a été suggéré que la répression de *cII* et *cIII* par Cro pourrait y jouer un rôle [15].

1.1.3 Maintien et sortie de la lysogénie

Boucle et coopérativité

CI, se fixant aux opérateurs O_{R1} et O_{L1}, inhibe fortement l'activité des promoteurs P_L et P_R, évitant ainsi la production de protéines qui pourraient déstabiliser l'hôte (protéines de recombinaison), voire le tuer (gènes *kil* et *S*, initiation de réplication à l'origine de réplication de Lambda dans *O*, médiée par *O* et *P*), ou inutiles (protéines formant la capside).

On voit ainsi qu'il est critique que la répression soit stable : sa levée, même brève, pourrait conduire à la dispartion de l'hôte, et à celle du prophage avec lui. Or un faible nombre de molécules sont impliquées (une molécule d'ADN, quelques dimères de CI) : on peut s'attendre à ce que les fluctuations thermiques dégagent, même brièvement, ces promoteurs.

Deux mécanismes stabilisent la lysogénie : la coopérativité de fixation de CI sur ses opérateurs et la formation d'une boucle d'ADN entre O_R et O_L (voir la figure 1.7). L'affinité d'un complexe CI_2 pour un opérateur, par exemple O_{R2}, est très fortement accrue par la présence des opérateurs O_{R1}, O_{L1} et O_{L2}, et CI forme jusqu'à des octamères sur ces quatre sites. Cette très forte coopérativité assure une forte occupation des sites, une fois un seuil de concentration de CI dépassé, renforçant ainsi la stabilité de la lysogénie [23, 24].

Notons qu'il existe un niveau supplémentaire de contrôle : CI active sa transcription quand il est fixé à O_{R2}, mais la réprime quand fixé en O_{R3} ; or l'affinité de CI_2 pour O_{R3} et O_{L3} est faible et leur occupation n'est pas facilitée par la présence de dimères de CI en O_{R2} ou O_{L2}, si bien que CI ne peut s'y fixer qu'en tirant partie de la coopérativité qui naît du rapprochement de ces deux sites lorque la boucle est formée. (voir la figure 1.7). Ainsi, au cours de la lysogénie, CI peut maintenir sa concentration en s'assurant que la répression ne sera pas levée (la boucle ne sera pas défaite).

On peut soulever ici deux questions. Les transcrits produits sous P_{RM} commencent immédiatement au codon *start* de *cI* ; la façon dont la traduction peut alors être initiée fait encore débat [25, 26] ; cela peut-il jouer un rôle dans la décision, le maintien de la lysogénie ou l'induction ? Deuxièmement, comment la lysogénie se maintient-elle lors du passage de la fourche de réplication de l'ADN, dont on peut s'attendre à ce qu'elle dégage les sites O_R et O_L ?

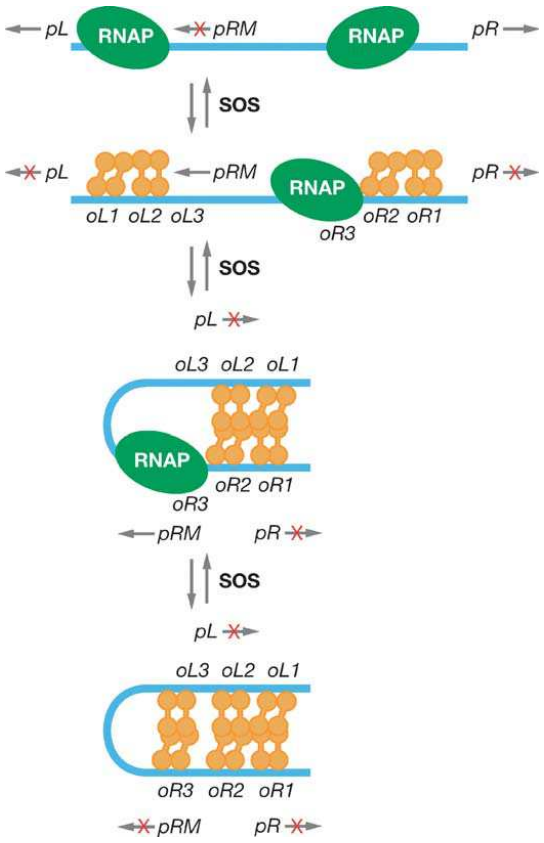


Fig. 1.7 – Maintien et sortie de lysogénie : en descendant, formation de la boucle d'ADN entre les sites O_L et O_R ; en remontant, sous l'effet du système SOS (clivage des dimères de CI par RecA), levée de la repression des promoteurs P_L et P_R et de l'activation de P_{RM}. En orange, les molécules individuelles de CI, qui se fixent sous forme de dimères coopérativement à leurs opérateurs ; RNAP : ARN-polymérase. Figure tirée de [11].

Induction de la lyse

Si la lysogénie est très stable, il existe pourtant un mécanisme de sortie : les protéines RecA de l'hôte clivent les dimères de CI, conduisant à une réponse coordonnée, telle que le prophage s'excise du chromosome, circularise puis suit un développement lytique (voir la figure 1.7). RecA est une protéine impliquée dans la réponse SOS de la bactérie à un endommagement de son chromosome. En laboratoire, l'induction peut ainsi être obtenue par exposition de lysogènes à un éclairage UV.

Le rôle de Cro pendant l'induction a été beaucoup discuté. Il s'avère que son action sur P_{RM} n'est pas essentielle, il permettrait plutôt, en réprimant l'expression de *cII* et *cIII* par son action inhibitrice sur P_R et P_L, d'éviter une réintégration dans le chromosome [27, 20, 28].

L'excision du prophage est réalisée par un complexe de protéines virales Int/Xis... alors que l'intégration est médiée par Int. La terminaison t_i est levée par N, uniquement pour les transcriptions initiées en P_L ; or le site *sib* forme une boucle sur le transcrit, dégradée par la RNaseIII (à nouveau), puis le transcrit est partiellement dégradé, empêchant la traduction de *int* et partiellement de *xis*[9]. CII active P_I, qui se trouve dans *xis*, et la transcription sous lui est terminée en t_i : seule Int est produite sous P_I.

Dans le lysogène cependant l'ADN viral est ouvert en *attP*, qui se trouve entre la fin du gène *int* et la terminaison t_i (et donc, surtout, entre *int* et *sib*), si bien que les transcrits produits sous P_R à la levée de la répression ne sont plus destabilisés et Int et Xis sont produits dans des proportions strictement identiques : le complexe Int/Xis est majoritairement formé, et l'excision peut se faire sans intégration ultérieure. On comprend ainsi d'autant mieux l'importance de l'absence d'activité de CII à l'induction [28].

On constaste ici encore la grande subtilité d'organisation des fonctions sur le génome et la diversité des processus à l'œuvre.

1.2 Approche dynamique

1.2.1 Un nouvel intérêt porté Lambda

Lambda est devenu un modèle pour l'étude de la dynamique d'expression génétique et sa régulation. J'en présente ci-dessous les principales raisons puis trois approches particulières.

Un système modèle pour l'étude de la dynamique

Le réseau décrit dans la section précédente est l'un des premiers à avoir été disséqué, beaucoup des processus évoqués y ont été observés pour la première fois. La bonne connaissance de sa structure génétique, des interactions protéiques impliquées en ont naturellement

9. [29] rapporte que la dégradation affecte essentiellement *int* ; [13] suggère que le transcrit est dégradé jusqu'en amont de *xis*, mais les travaux cités sont antérieurs à [29].

fait le principal système, avec l'opéron Lac, auxquels sont confrontés les modèles théoriques. De plus, une étude expérimentale sur Lambda sera peu sujette, comparé à d'autres systèmes, à des incertitudes sur les facteurs intervenant dans les phénomènes observés.

Il remplit une fonction simple de « choix » entre deux voies de développement, et permet de plus le saut d'un état stable à un autre : il a ainsi suscité de nombreux travaux sur la multistabilité des réseaux génétiques. Notons que cette question est particulièrement importante dans l'étude du développement des organismes pluricellulaires.

Nous l'avons vu, une grande souplesse (biais dans le choix) est obtenue par la diversité à la fois des signaux intégrés et des processus mis en jeu. En particulier, deux logiques simples et complémentaires sont à l'œuvre : commutateur exclusif Cro/CI et voie de signalisation par CII. Avec des précautions, on peut de plus considérer ce réseau comme très « optimisé ». On constate une organisation du génome poussée, très modulaire et parfois étonnamment subtile, avec peu de redondances ; la diversité des processus et des organisations évoquée apparaissent alors comme un moyen de « tirer parti » des possibilités offertes. Tout cela est le produit d'une évolution « accélérée » par la facilité des transferts horizontaux et de la pression forte exercée par la compétition avec d'autres phages, avec *E. coli*, et par la contrainte subie par le génome de devoir entrer dans une capside.

Enfin, l'infection d'une cellule par un virus peut être vue comme un cas aussi proche que possible de transitoire d'un état initial « minimal » bien défini (au moment de l'infection, seule une ou quelques molécules d'ADN viral est présente, sans aucune protéine de Lambda) vers l'un ou l'autre de deux états stationnaires possibles ; avec cette réserve que l'état de l'hôte joue une rôle important.

Agencement temporelle des cascades lytiques et lysogéniques

Nous l'avons vu, il est difficile de décrire dans une succession temporelle les cascades de régulations, alors même que leur agencement temporel joue un rôle déterminant.

Utilisant des plasmides portant des gènes codant pour la protéine fluorescente EGFP sous P_{RE} ou sous $P_{R'} - t_{R'}$, les auteurs de [15] ont pu mesurer les activités de CII et Q au cours de l'infection d'une population de bactéries par différentes souches de phages. Ils ont notamment trouvé un délai important dans l'activité de Q par rapport à celle de CII, bien qu'elles soient produites par traduction du même transcrit.

Une méthode équivalente a été utilisé dans deux études sur l'induction de la lyse : dans [30] le gène *egfp* est derrière P_R ou $P_{R'} - t_{R'}$, dans [28] il est derrière P_R, P_{RE} ou $P_{R'} - t_{R'}$.

Rôle des fluctuations stochastiques

Lambda fut le premier modèle utilisé dans des simulations stochastiques de la dynamique d'expression d'un réseau de régulation génétique [22]. Depuis, de nombreux modèles d'expression stochastique mettent en avant le rôle que les fluctuations pourraient jouer dans la décision lyse/lysogénie [31, 32, 33, 34].

D'importantes fluctuations dans les activités de CI et Q ont été observées après induction de la lyse par irradiation UV [30]. Les auteurs de [16], trouvant une forte corrélation entre le résultat de l'infection et la taille de la bactérie au moment de l'infection, ont suggéré qu'elles joueraient un rôle restreint ; ils ne mesurent cependant pas l'expression de gènes de la région d'immunité.

Il apparaît ainsi que la décision est un phénomène aléatoire fortement biaisé par l'état de l'hôte, mais aucune étude expérimentale n'a encore permis d'estimer l'importance relative des aspects « aléatoire » et « déterministe ».

Stabilité de la lysogénie et induction

La stabilité de la lysogénie[10] est particulièrement étonnante dans un contexte où l'on s'attend à de fortes fluctuations stochastiques.

Dans [35], les auteurs ont remplacé une portion du génome de Lambda (dans *rexA* et *rexB*, juste derrière *cI*) par *gfp* de façon à mesurer l'expression de *cI* au cours de la lysogénie sans modifier le nombre de bases entre O_R et O_L. Ils ont trouvé une grande variation de concentration de répresseurs de cellule à cellule ; par ailleurs, ils ont montré que la plupart des phages produits par induction spontanée sont mutés dans *cI* ou P_{RM}.

Deux études récentes ont confirmé le rôle de la boucle formée entre O_R et O_L dans le maintien de la lysogénie [36, 24].

1.2.2 Pour une description plus complète

Le fonctionnement du réseau de régulation de Lambda est encore très discuté. Les trois études expérimentales sur sa dynamique évoquées plus haut [15, 30, 28] ont plusieurs inconvénients : les mesures des première et troisième ont été faites sur des populations entières, masquant les variations d'une cellule à l'autre ; l'utilisation d'un plasmide rapporteur ne tient pas compte des effets liés à l'organisation sur le génome et pour la deuxième ajoutent des fluctuations indépendantes du système étudié (nombre de copies de plasmide, transcriptions des gènes viraux et des gènes rapporteurs séparées) ; les activités des différentes protéines régulatrices sont mesurées dans des populations différentes, ce qui ne permet pas d'estimer leurs interactions ; leur activité est mesurée, et non leur production.

Ainsi, la démarche suivie au cours du travail exposé dans cette thèse et décrite en détail au chapitre suivant a été la suivante :
1. pour obtenir une mesure de la dynamique de l'activité génétique, on utilisera des protéines fluorescentes produites en conjonction avec les protéines virales qui nous intéressent ;
2. pour tenir compte le plus possible des effets physiques liés à la position des gènes, promoteurs et terminaisons sur le génomes (délais induits par les terminaisons de transcription,

10. La fréquence d'induction spontanée dans des souches *recA* est inférieure à 2.10^9 par cellule par génération, soit moins que le taux de mutation des gènes impliqués [31].

sites *nut* d'utilisation de N, gêne de polymérases transcrivant en sens inverse voire levée de la régulation par les polymérases passant sur un opérateur [37], formation d'une boucle d'ADN), les gènes codant pour des protéines fluorescentes seront insérés dans le génome de Lambda ;

3. pour mesurer directement les corrélations d'expression, deux gènes codant pour des protéines fluorescentes de couleurs différentes seront insérés à la fois ;

4. pour une description plus riche du réseau, plusieurs couples de lieux d'insertion seront retenus, et plusieurs constructions réalisées ;

5. enfin, des mesures à différents nombre de copies du réseau seront faites.

Chapitre 2

Étude expérimentale

2.1 Principe

Lors de la lysogénie, *cI* est le seul gène intervenant dans la décision lyse/lysogénie exprimé. Il inhibe fortement l'expression des autres gènes. Partant d'un lysogène muté de façon à rendre CI thermosensible, on peut en principe, par un choc de température, retrouver une situation similaire à l'infection de l'hôte par Lambda : le génome de Lambda est présent dans la cellule mais aucune de ses protéines n'y est active. On s'attend à ce qu'une des deux voies soit suivie une vingtaine de minutes après le choc de température, comme c'est le cas lors d'une infection [15].

En insérant des gènes codant pour des protéines fluorescentes dans l'ADN viral, on peut alors mesurer l'activité transcriptionnelle du réseau au cours de l'établissement de la lyse ou de la lysogénie, et étudier la stabilité de ces deux états.

2.2 Constructions génétiques

2.2.1 Le système étudié

Portion du génome de Lambda

J'ai gardé toute la portion du génome de Lambda contenant les éléments intervenant dans la décision lyse/lysogénie : du gène *cIII* au promoteur P_{OOP} (bases 33299 à 38599). De cette séquence, seuls les gènes *rexA*, *rexB*, *SieB*, *ral* et *Ea10* n'interviennent pas dans la décision [1]. Des séquences non-codantes interviennent cependant en *cis* dans la régulation (notamment lors de l'autorégulation de *N*) et les délais de transcription peuvent jouer un rôle dans l'établissement de l'une des deux voies : il m'a semblé important de ne pas amputer cette portion du génome de Lambda.

Les séquences au-delà de *cIII* et P_{OOP} n'ont cependant pas pu être conservées. Dans le premier cas, le produit du gène *kil* aurait tué l'hôte, dans le deuxième cas l'origine de

1. Les trois premiers protègent le lysogène de l'infection par d'autres bactériophages.

réplication *ori* et le gène *O* impliqué dans la réplication de l'ADN viral auraient fortement perturbé le système.

L'allèle *cI-ts2*

La mutation ts2 a été introduite dans le gène *cI* [38, 39] : le codon 224 y est GAA, au lieu de AAA dans le gène sauvage. Son produit possède les mêmes propriétés que celui de l'allèle sauvage (formation de multimères, fixation aux sites O_R et O_L, clivage par RecA), mais il est de plus thermosensible : il se dénature à haute température et ne se renature en principe pas. « En principe » car si je n'ai pas trouvé d'étude qui le montre, [40] indique que les mutations U32 et U37 de *cI*, C-terminales (« de type A » dans la terminologie de cette étude) comme ts2, conduisent à des répresseurs inactifs après passage à haute température et retour à une température permissive, alors que les mutations t1 et 857, N-terminales (« de type B »), conduisent à des répresseurs qui restent actifs. De plus, les résultats de la troisième expérience présentée au paragraphe 2.3.3 et sur la figure 2.7 vont dans ce sens.

À noter que le répresseur CI-ts2 possède une affinité à l'ADN plus faible et forme plus difficilement des dimères et des octamères que le variant sauvage, est stable à 30 °C mais instable à 35 °C et que la lyse est facilement induite par exposition à un rayonnement UV [39] : il est « moins efficace » que le variant sauvage.

Dans [28], un choc de température est utilisé pour induire la lyse chez des lysogènes possédant l'allèle *cI857* et suivre la dynamique de sortie de lysogénie, mais une possible renaturation du répresseur n'est pas discutée.

Terminaisons de transcription

Les terminaisons de transcription *TeT7* et *rrnBT1* [41] ont été ajoutées de part et d'autre de la portion du génome de Lambda conservée, de façon à éviter les transcriptions d'ADN viral initiées ailleurs, ou que les transcriptions initiées à un promoteur de Lambda ne se poursuivent au-delà.

2.2.2 Lieux d'insertion des gènes rapporteurs

Trois constructions ont été réalisées, permettant de suivre l'activité de trois couples d'éléments (gènes, promoteurs, terminaisons) : les gènes *egfp* et *mCherry* ont été insérés dans le génome de Lambda, de façon à être transcrits sur les mêmes ARN qu'un couple de gènes choisis, ou plus généralement à rapporter l'activité transcriptionnelle au *locus* où ils ont été insérés. Ces constructions sont représentées schématiquement sur la figure 2.1.

Dans la première, les gènes *cI* et *cro* ont été marqués, ce qui doit permettre de voir le résultat de la décision : lors de la lysogénie, *cI* est exprimé alors que *cro* est inhibé, et inversement lors de la lyse. Elle doit aussi permettre d'apprécier le rôle de la compétion entre CI et Cro pour la fixation sur O_R dans l'établissement de l'une des deux voies. Cette

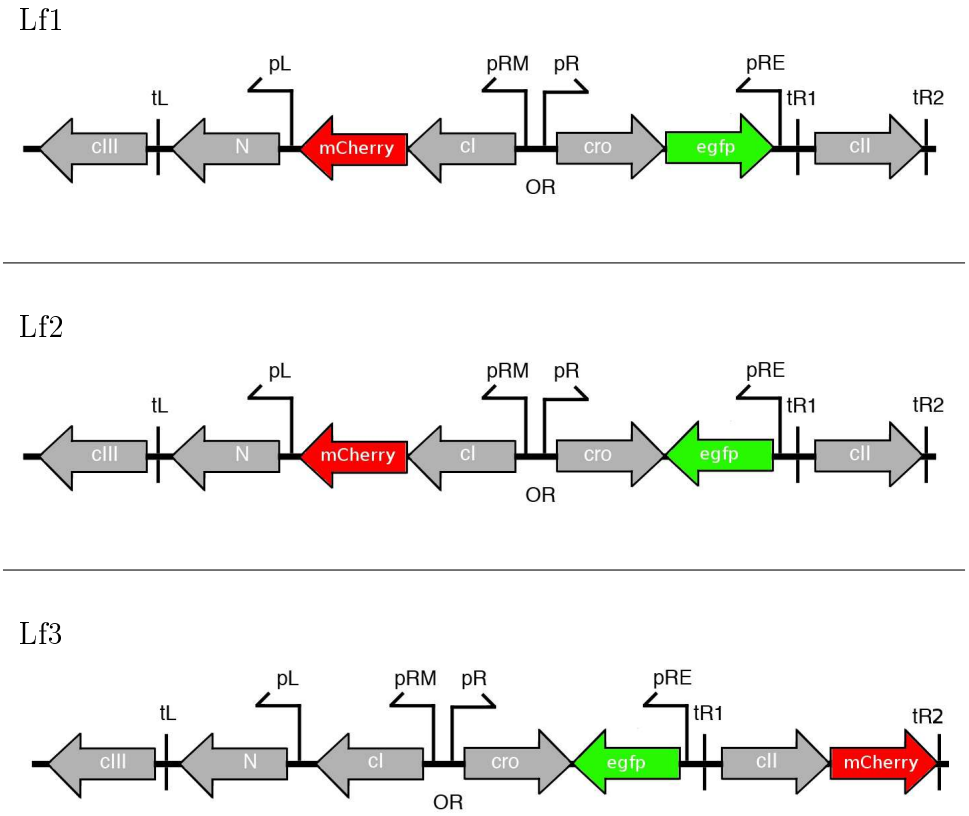


Fig. 2.1 – Schémas indiquant les lieux d'insertion des gènes codant pour des protéines fluorescentes (*mCherry* et *egfp*) et leur orientation dans les trois constructions réalisées.

construction sera notée « Lf1 » dans la suite ; c'est la seule à avoir été utilisée dans cette étude.

Dans la deuxième (« Lf2 »), le gène *cI* est marqué et le gène *egfp* est inséré directement sous le promoteur *pRE* : on doit ainsi pouvoir suivre l'établissement de la lysogénie, et voir d'éventuelles « tentatives avortées ».

Enfin, la dernière construction (« Lf3 »), dans laquelle *cII* est marqué et le gène *egfp* inséré directement sous le promoteur *pRE*, doit permettre de mieux comprendre le rôle de CII en suivant à la fois sa production et son activité [2].

Bien d'autres couples pourraient être choisis, notamment *N* et *cIII* mériteraient d'être marqués, mais les trois couples présentés m'ont semblé être les plus pertinents. D'une construction à la suivante, un même gène codant pour une protéine fluorescente a été inséré au même endroit : on pourra ainsi comparer aisément les mesures faites sur ces trois constructions.

Deux autres techniques pourraient être envisagées pour suivre la dynamique de ce réseau : la fusion de protéines de Lambda avec des protéines fluorescentes et le clonage dans un plasmide d'un promoteur de Lambda (et éventuellement d'une terminaison de transcription) suivi d'un gène codant pour une protéine fluorescente. La première technique, si elle plus « fidèle » (pas de bruits supplémentaire de traduction) comporte le risque de perturber l'action des protéines ainsi marquées et de ne pas conduire à des niveaux de fluorescence détectables. La deuxième technique a été évoquée à la fin du chapitre précédent et est utilisée par exemple par les auteurs de [15, 30, 28] pour étudier ce réseau de Lambda. Remarquons qu'elle a l'avantage de le laisser intact.

2.2.3 Fusion transcriptionnelle de gènes codant pour des protéines fluorescentes

Le choix de *egfp* et *mCherry*

Les pics des spectres d'excitation et d'émission des protéines fluorescentes choisies doivent être, d'une protéine à l'autre, aussi éloignés que possible [3]. Les couples de protéines possibles sont ainsi « bleu-jaune (ou rouge) » et « vert-rouge ». L'utilisation de protéines bleues, telle que CFP, n'est cependant pas recommandée : la lumière d'excitation, dans le bleu profond, est toxique pour les bactéries. Elles ne se prêtent donc pas à un éclairage

2. Une difficulté se pose dans le marquage de *cII* : son transcrit est déstabiliséé par un court ARN antisens produit sous le promoteur *pOOP*, lequel se trouve immédiatement après le gène. En insérant *mCherry* au-delà de P_{OOP}, on ignorerait ce niveau de régulation. J'ai préféré l'insérer immédiatement après le codon stop de *cII*, au risque de perturber l'action de P_{OOP} sur *cII*.

3. Dans la suite, par commodité, je parlerai de couleur de fluorescence ou de protéine : ainsi, « protéine rouge » désignera une protéine qui réémet principalement dans le rouge et « fluorescence rouge » la lumière correspondante.

fréquent, comme ce doit être le cas dans des mesures de dynamique d'expression génétique. La protéine EGFP est à la fois la plus brillante et la plus photostable des protéines vertes ; deux protéines rouges sont brillantes et assez photostables : mCherry et TagRFP-T, récemment inventée [42]. mCherry est légèrement moins brillante et stable que TagRFP-T, mais elle est un peu plus « éloignée » de EGFP et surtout elle a un temps de maturation beaucoup plus court [42].

En effet, pour déterminer aisément leur taux de production, les protéines fluorescentes choisies doivent aussi avoir des temps de maturation courts, d'une demi-heure ou moins. mCherry a un temps de « demi-maturation » d'environ 15 min à 37 ° C ([43], Informations Supplémentaires Tableau 1). La référence [43] ne donne pas d'estimation pour EGFP ; un temps moyen de maturation de 6,5 min pour la GFP a été mesuré par les auteurs de [44], mais il dépend fortement de la variante de la protéine considérée [45] ; une valeur de 15 min est communément admise (elle est assumée telle dans [15] par exemple).

Il serait utile de déterminer les temps relatifs de maturation de mCherry et EGFP dans les conditions de cette étude.

Séquence de fixation de ribosome

Les gènes *egfp* et *mCherry* ont été insérés dans le génome de Lambda aux *loci* indiqués au paragraphe 2.2.2, juste après les codons stop des gènes *cI*, *cro* ou *cII*, précédés d'une séquence de fixation de ribosome censée conduire à de forts taux de traduction ([46], Informations Supplémentaires) : TAAGGAGGAAAAAAA.

2.2.4 Les vecteurs

Les constructions ont été clonées dans les plasmides pZC320 [47] et pMK (pCR4 pour Lf2), possédant respectivement les origines de réplication *ori-2* de F et ColE1. Le premier a un nombre de copies par chromosome de 1/2 environ à 30 ° C dans du LB, soit un nombre copies absolu quasiment de 1, le deuxième de plusieurs dizaines. On pourra ainsi qualitativement estimer l'effet du nombre de copies du génome viral sur les processus de décision.

2.3 Contrôles et mesures préliminaires

2.3.1 Extinction de fluorescence (*Photobleaching*)

Éclairées à leur longueur d'onde d'absorption, la plupart des protéines fluorescentes subissent une transition vers un état non fluorescent. Il est nécessaire de tenir compte de cet effet si l'on veut relier les niveaux de fluorescence mesurés aux taux de production de EGFP et mCherry.

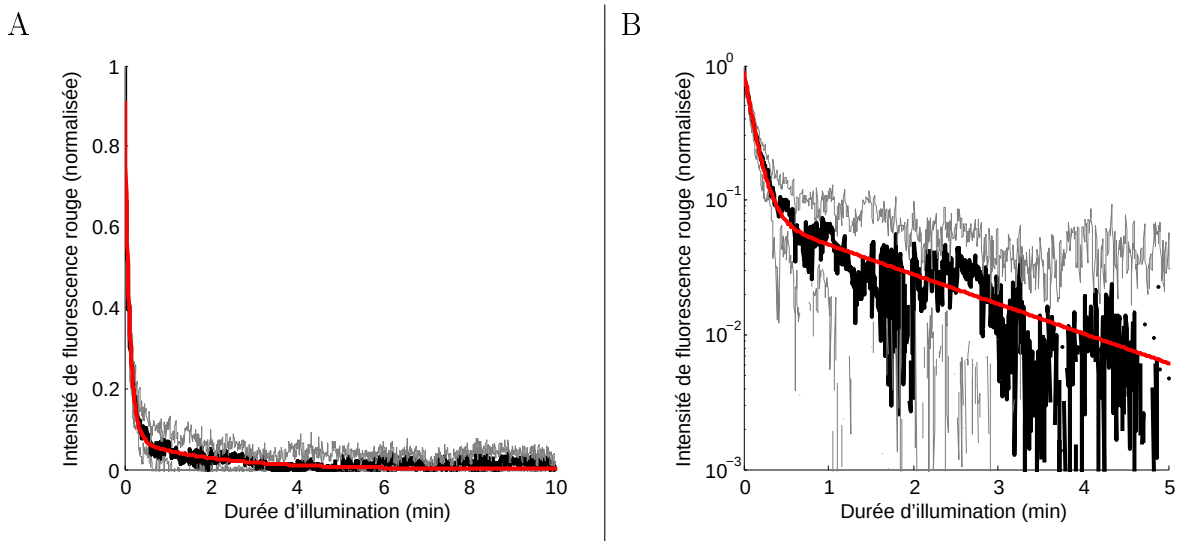


Fig. 2.2 – Moyenne (courbe noire) et moyenne plus ou moins une déviation standard (courbes grises), sur cinq expériences, de l'intensité de fluorescence rouge de bactéries pZC-Lf1 individuelles, sous éclairage continu. La valeur du fond, à la fin de l'expérience, a été soustraite puis l'intensité divisée par sa valeur initiale. Courbe rouge : ajustement obtenu par une somme de deux exponentielles : $I(t) = Ae^{\lambda_+ t} + Be^{\lambda_- t}$, de paramètres $A = 0,83$, $B = 0,07$, $\lambda_+ = -0.15\ s^{-1}$ et $\lambda_- = -8.1 \times 10^{-3}\ s^{-1}$. (A) échelles linéaires, (B) échelles log-lin.

La courbe d'extinction de fluorescence de mCherry a été obtenue en éclairant de façon continue des bactéries l'ayant exprimée et en mesurant l'intensité de fluorescence [4]. Le meilleur ajustement a été obtenu par une somme de deux exponentielles (figure 2.2 A). La décroissance initiale rapide est bien visible sur la courbe en échelles log-lin (figure 2.2 B). Cela suggère que mCherry peut subir deux transitions, l'une rapide et réversible, l'autre lente et irréversible [5]. De tels processus sont décrits dans [48] à propos d'une autre protéine fluorescente, ECFP, variante de GFP. Dans [43], une décroissance initiale rapide

4. Des bactéries pZC-Lf1 ont poussé jusqu'à saturation, puis elles ont été lavées et resuspendues dans du PBS (voir en annexe A.2.2)) supplémenté en chloramphénicol, de manière à bloquer la production de mCherry au cours de la mesure.

5. Notons P_n la protéine dans son état natif, et P_1 et P_2 les produits non-fluorescents. Considérons que P_n participe aux deux réactions :

$$P_n \underset{k_2}{\overset{k_1}{\rightleftarrows}} P_1 \quad \text{et} \quad P_n \xrightarrow{k_3} P_2$$

À l'instant initial, toutes les protéines sont dans leur état natif. En notant I l'intensité de fluorescence (proportionnelle au nombre de protéines P_n), normalisée de telle sorte que $I(0) = 1$, il vient :

$$I(t) = Ae^{\lambda_+ t} + Be^{\lambda_- t}$$

avec $\lambda_\pm = \frac{1}{2}\left(-(k_1+k_2+k_3) \pm \sqrt{(k_1+k_2+k_3)^2 - 4k_2k_3}\right)$, $A = -\frac{k_1+k_3+\lambda_-}{\lambda_+-\lambda_-}$ et $B = 1 - A$.

Remarque : les taux k_i dépendent de l'intensité d'éclairage, linéairement dans le cas de la protéine ECFP [48].

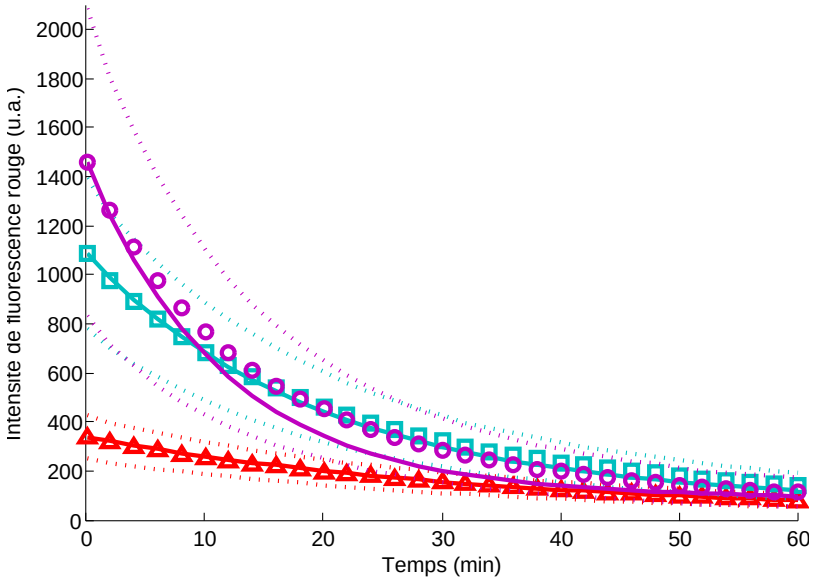


Fig. 2.3 – Intensité de fluorescence rouge de bactéries pZC-Lf1 individuelles sous éclairage intermittent. Un intervalle de 2 min sans éclairage sépare deux mesures, de durées d'illumination de 0,4 (triangles rouges), 0,7 (carrés cyan) et 1,2 s (ronds magenta). Courbes pleines : intensités prédites, dans ces conditions, par l'ajustement des données de l'expérience en éclairage continu. Courbes pointillées : valeurs expérimentales plus ou moins une déviation standard. Une durée de 0,2 s a été ajoutée à la durée d'illumination (les valeurs indiquées en tiennent compte). Un recouvrement de la fluorescence dans le noir se traduirait ici par une fluorescence mesurée plus importante que la valeur prédite par l'ajustement.

d'une autre variante de EGFP est aussi rapportée, par contre la courbe d'extinction de mCherry y est quasi-exponentielle (Informations Supplémentaires, Figure 1). Des mesures sous éclairage intermittent semblent indiquer qu'il n'y a pas de recouvrement de la forme native de mCherry dans le noir (figure 2.3). Dans [42] cependant, les auteurs indiquent un taux de recouvrevement de mCherry après une à deux minutes dans le noir de 14% (dans des conditions différentes : une intensité d'éclairage plus grande et mCherry fusionnée à une protéine dans une cellule eukaryote). J'utiliserai les mesures que j'ai réalisées pour corriger les niveaux de fluorescence mesurés dans des conditions identiques.

Dans les conditions d'éclairage utilisées par la suite, je n'ai pas détecté d'extinction de fluorescence de EGFP.

2.3.2 Phototoxicité

La lumière, en particulier dans les courtes longueurs d'onde, ou les produits d'extinction de fluorescence peuvent être toxiques pour les bactéries. L'intensité et la durée d'excitation des protéines fluorescentes (décrites en annexe A.3.2) ont été choisies suffisamment faibles

	Temps de division (min)	Vitesse de croissance (u.a.)
TOP10 sans éclairage	21 (6)	1,5 (0,3)
pZC-Lf1 avec éclairage	26 (12)	1,3 (0,5)

Tab. 2.1 – Croissance de bactéries n'exprimant pas de protéines fluorescentes (TOP10) sans éclairage de fluorescence et de bactéries de fond génétique identique mais portant la construction pZC-Lf1, sous les conditions d'éclairage décrites au paragraphe A.3.2. Les bactéries ont poussé sur du LB-agar et l'échantillon a été maintenu à 37 ° C. Sont indiquées les moyennes et déviations standard (entre parenthèses) du temps de division (durée séparant la naissance de la division d'une bactérie) et de sa vitesse de croissance, définie comme le rapport de son augmentation de longueur (en unités arbitraires) au cours de son cycle et de son temps de division. Elles ont été obtenues par des mesures sur 312 bactéries TOP10 (issues de 7 bactéries, sur une expérience) et 377 bactéries pZC-Lf1 (issues de 12, sur trois expériences). À noter que l'intervalle de 5 min séparant deux images et la difficulté d'estimer si une cellule est sur le point de se diviser ou si elle vient de le faire induisent une incertitude de l'ordre de 5 min sur le temps de division.

pour peu influer sur le temps de division et le taux de croissance des bactéries exprimant ces protéines (tableau 2.1), mais suffisantes pour qu'on puisse clairement distinguer les deux états, lyse et lysogénie, des bactéries pZC-Lf1 (voir le paragraphe suivant et la figure 2.4).

2.3.3 Levée de la répression

Pour qu'on puisse observer l'établissement d'une des voies, lyse ou lysogénie, il faut incuber des lysogènes à haute température suffisamment longtemps pour dénaturer les répresseurs CI, mais un pendant temps suffisamment court pour ne pas forcer la voie lytique et limiter le choc de température.

Une première expérience montre qu'il est possible de distinguer deux états, interprétés comme « lysogénique » et « lytique », et de faire basculer une population entière de l'un à l'autre. Considérant la souche pZC-Lf1, on s'attend à ce que dans l'état lysogénique, CI (et donc mCherry) soit produite et la production de Cro (et donc EGFP) fortement réprimée ; inversement, dans l'état lytique, la production de mCherry doit rester faible alors que EGFP doit être fortement produite. Une colonie de pZC-Lf1 exprimant *mCherry* et n'exprimant pas *egfp* a été inoculée dans du milieu minimal (MM, voir en annexe A.2.2) et incubée pendant une nuit à 30 ° C. Cette culture a ensuite été inoculée dans du MM et incubée une nuit à 30 ° C d'un côté et dans du LB et incubée une nuit à 37 ° C de l'autre. La figure 2.4 présente la fluorescence mesurée sur chacune des cultures ainsi obtenues. Elles s'y distinguent bien l'une de l'autre. Un milieu pauvre tel que le MM est connu pour favoriser la lysogénie, alors qu'un faible nombre de copies du génome et un milieu riche tel que le LB favorisent la lyse. À 37 ° C, CI-ts2 a perdu beaucoup de son activité, ce qui, dans ces conditions défavorables, empêche la lysogénie d'être maintenue.

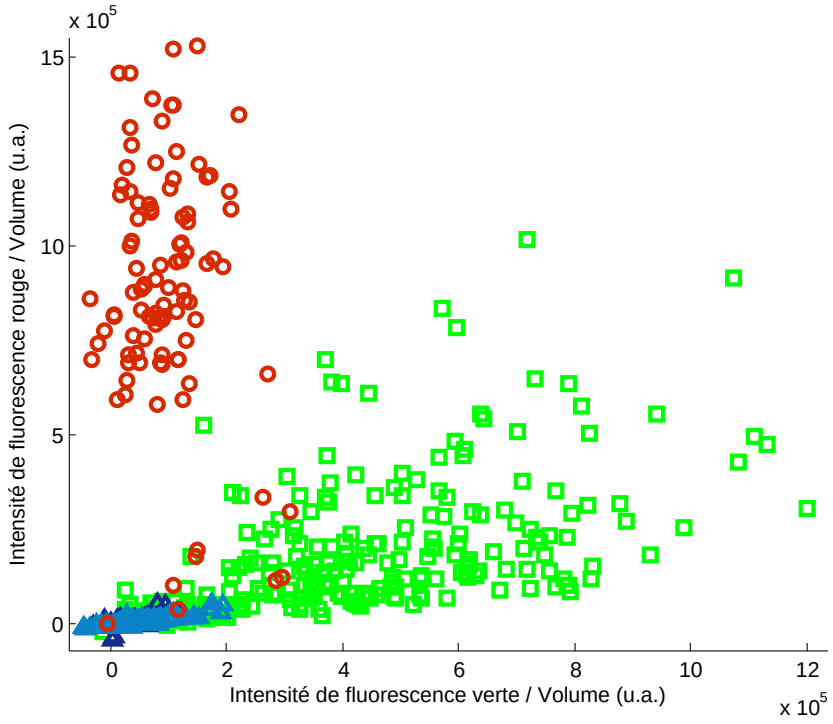


Fig. 2.4 – Diagramme de fluorescence de bactéries pZC-Lf1 individuelles après une nuit à 30 ° C dans du MM (cercles rouges) et après une nuit à 37 ° C dans du LB (carrés verts). Pour référence, sont aussi indiquées des bactéries TOP10 ayant poussé à 30 ° C dans du MM (triangles bleu foncé) et à 37 ° C dans du LB (triangles bleu clair) : les niveaux de fluorescence des états « lysogénique » et « lytique » se distinguent bien de l'autofluorescence. u.a. : unités arbitraires.

Ainsi, on peut bien interpréter les deux groupes visibles sur la figure 2.4 comme une population constituée majoritairement de lysogènes (cercles rouges) et une population de bactéries ayant suivi la voie lytique (carrés verts).

Il s'agit maintenant de déterminer le temps minimal à passer à haute température pour lever la répression. Cette température a été fixée à 42 ° C : si le choc est violent pour la bactérie, une température élevée permet cependant en principe d'en limiter la durée. Des bactéries pZC-Lf1 lysogènes ont été mises à pousser dans du MM à 30 ° C jusqu'à une OD de 0,2 ; elles ont été lavées, resuspendues dans du LB et laissées un temps t à 42 ° C ; elles ont enfin été lavées et resuspendues dans du PBS supplémenté en chloramphénicol. La figure 2.5 montre les niveaux de fluorescence mesurés pour $t = 0, 20, 40, 60, 80, 100$ et 120 minutes.

Pour $t = 0$ min, on retrouve bien une population de fluorescence rouge élevée et de fluorescence verte faible ; pour $t = 120$ min, on peut considérer que toutes les bactéries ont basculé vers l'état « lytique ». Pour des temps intermédiaires, on voit la population se déplacer progressivement du rouge vers le vert. On aurait pu s'attendre à voir l'état

« lysogénique » se dépeupler au profit de l'état « lytique » ; rien ne garantit cependant qu'un état stationnaire ait été atteint, et il n'est pas surprenant que, pendant un temps, CI et Cro soient produites simultanément. De plus, mCherry n'étant pas dégradée, seule la dilution fait baisser le niveau de fluorescence rouge. Cela peut expliquer en particulier qu'au bout de 20 min, les bactéries fluorescent davantage dans le vert tout en gardant un niveau de fluorescence dans le rouge comparable aux lysogènes.

On peut avancer deux arguments supplémentaires pour expliquer l'évolution progressive d'un état vers l'autre au cours du passage à 42 ° C : la construction est portée par un plasmide dont le nombre de copies varie typiquement de 1 à 2, on peut imaginer que pendant un temps deux copies dans deux états différents coexistent dans une cellule ; deuxièment, [16] décrit une voie de développement mixte lyse et lysogénie dont on voit peut-être la trace ici.

Prenant comme référence les groupes visibles sur la figure 2.4 et les groupes de bactéries à $t = 0$ min et $t = 120$min, j'ai délimité à la main deux régions du diagramme rouge-vert (figure 2.5 A) ; la figure 2.5 B montre la proportion de bactéries dans chaque région en fonction du temps passé à haute température. Une grande majorité de bactéries semble ainsi avoir basculé au bout de 60 min. Les mesures ont cependant été faites juste après le choc. Les protéines fluorescentes n'acquérant leur forme native que plusieurs minutes après avoir été produites[6], les niveaux de fluorescence mesurés à t correspondent à des protéines produites plus tôt.

La même expérience a été répétée, avec cette fois un passage à haute température dans du MM au lieu du LB. Les résultats sont présentés sur la figure 2.6. On ne peut plus maintenant distinguer deux groupes. La fluorescence verte n'augmente pas avec le temps passé à 42 ° C ; la fluorescence rouge diminue, peut-être parce que CI ne peut plus activer sa propre production sous pRM, celle-ci n'étant plus initiée qu'au promoteur pRE, moins fort : la lysogénie se maintient mal, mais l'état stationnaire « lytique » n'est pas atteint.

Pour s'affranchir du problème du temps de maturation des protéines et séparer l'étape de dénaturation des répresseurs du choix de la voie suivie, une expérience légèrement différente a été réalisée : les bactéries subissent le choc thermique dans du PBS, puis sont lavées et resuspendues dans du LB, et incubées pendant 2h à 30 ° C. Elles sont alors lavées et resuspendues dans du PBS supplémenté en chloramphénicol. La figure 2.7 présente les niveaux de fluorescence mesurés. Procédant à la même analyse que précédemment, on constate maintenant qu'elles ont pour l'essentiel basculé au bout de 30 min. Ici la question des protéines mCherry initialement présentes ne se pose pas : elles ont été largement diluées pendant l'induction de 2h (les bactéries s'y divisient trois à quatre fois). Remarquons que si les répresseurs se renaturaient, on s'attendrait à ce qu'une fraction de bactéries lysogènes soit trouvée après les deux heures d'incubation à 30 ° C.

6. La température peut avoir un effet important sur le temps de maturation, celui-ci croissant avec celle-la [45].

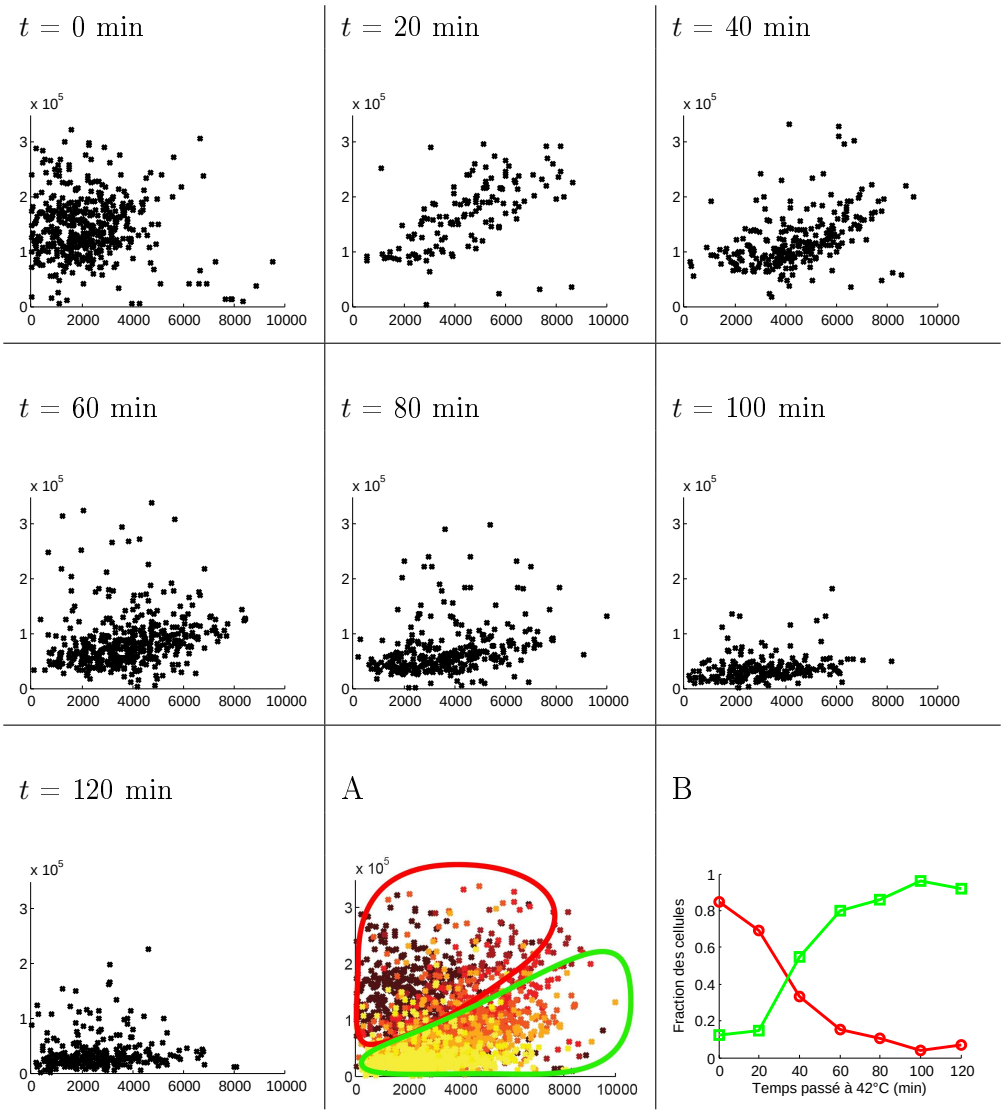


Fig. 2.5 – Levée de la répression : diagrammes de fluorescence de bactéries pZC-Lf1 ayant passé un temps t dans du LB à 42 ° C, pour t variant de 0 à 120 min. Les diagrammes sont superposés en (A), avec une couleur variant du noir ($t = 0$ min) au jaune ($t = 120$ min) ; deux régions peuvent être dessinées : « lysogénique » (contour rouge : fluorescence rouge élevée, fluorescence verte faible) et « lytique » (contour vert). (B) Fraction des bactéries dans les régions « lysogénie » (cercles rouges) et « lyse » de (A) (carrés verts). Les huit diagrammes : en abscisse (resp. ordonnée), intensité de fluorescence verte (resp. rouge) divisée par le volume de bactéries individuelles, en unités arbitraires.

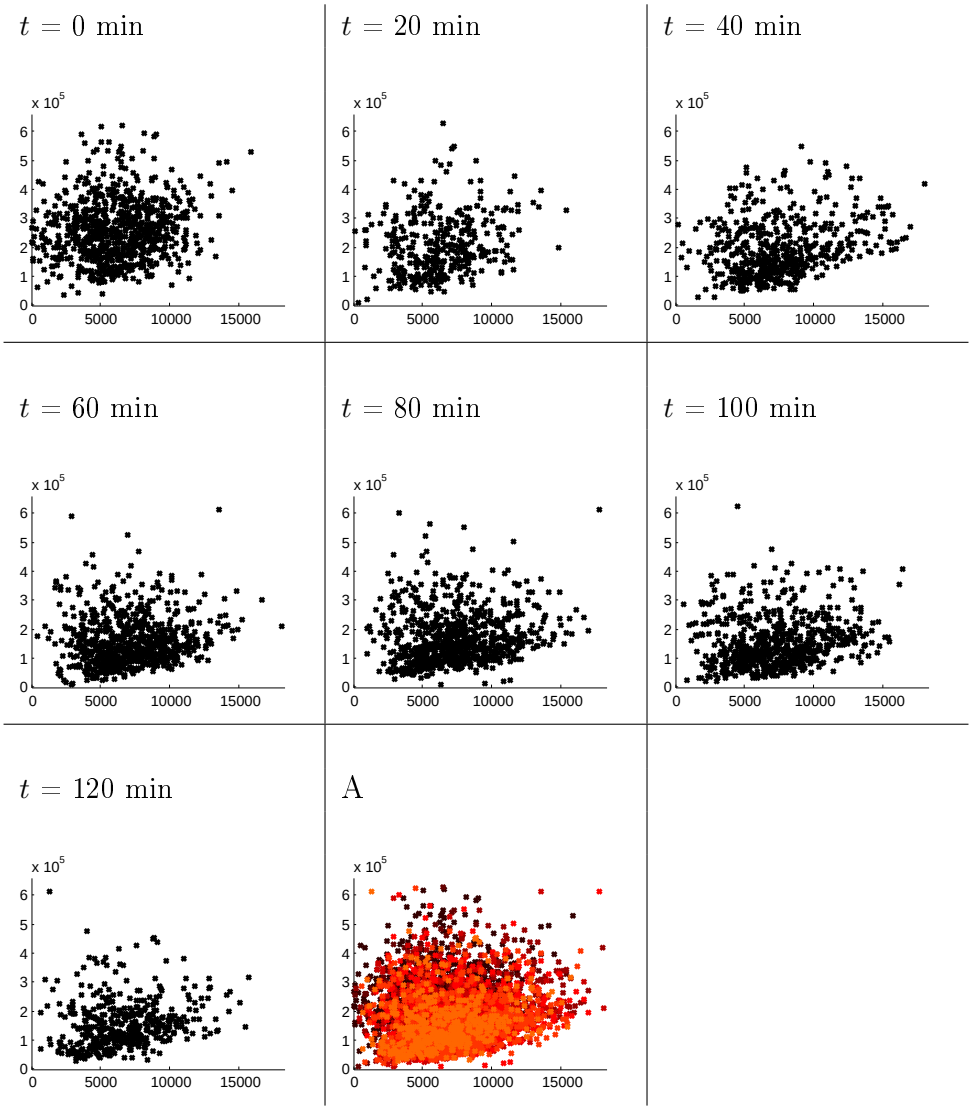


Fig. 2.6 – Légende identique à la légende de la figure 2.5, pour des bactéries ayant passé un temps t à 42 ° C dans du MM.

En conclusion, un passage de 30 min à 42 ° C semble permettre de lever complètement la répression sans forcer l'issue de la décision. Il serait bon cependant de tester la sensibilité du système au temps précis du choc et à sa température : soit en se plaçant dans des conditions où les probabilités de suivre chacune des deux voies sont comparables (telles que décrites dans [16] par exemple) et en estimant la dépendance des proportions de bactéries dans chaque état aux paramètres du choc ; soit en mesurant la dépendance à ces paramètres des taux d'expression de chaque gène codant pour une protéine fluorescente.

[28] suggère qu'un choc de 5 min est suffisant à l'inactivation complète des répresseurs.

Remarque : des réglages d'éclairage de fluorescence différents ont été utilisés lors des mesures présentées dans les figures 2.4, 2.5, 2.6 et 2.7 : on ne peut pas comparer d'une expérience à l'autre les valeurs de fluorescence mesurées.

2.4 Résultats

Cette section présente les résultats de mesures de dynamique d'expression, dans différentes conditions, des gènes codant pour des protéines fluorescentes de la construction Lf1 (*mCherry* derrière *cI* et *egfp* derrière *cro*). J'ai considéré, pour chaque bactérie, l'intensité de fluorescence de chaque couleur divisée par le volume, en fonction du temps : on s'attend en effet à ce que, CI et Cro étant des facteurs de transcription, ce soit leur concentration qui soit régulée[7]. Les intensités de fluorescence ont été corrigées de manière à ce qu'elles ne reflètent que les protéines produites au cours de la mesure ; l'extinction de fluorescence de mCherry a aussi été corrigée. Ces corrections sont détaillées en annexe A.4.2. Les grandeurs considérées sont ainsi, à un facteur multiplicatif près, les concentrations de protéines fluorescentes produites depuis le début de la mesure.

Les autocorrélations et corrélations croisées des intensités de fluorescence corrigées divisées par le volume sont aussi présentées. La corrélation de deux variables X et Y à deux temps t_1 et t_2 est définie ici par :

$$C(X,t_1;Y,t_2) = \frac{cov(X(t_1),Y(t_2))}{\sigma_{X(t_1)}\sigma_{Y(t_2)}} = \frac{\langle X(t_1)Y(t_2)\rangle - \langle X(t_1)\rangle\langle Y(t_2)\rangle}{\sqrt{\langle X(t_1)^2\rangle - \langle X(t_1)\rangle^2}\sqrt{\langle Y(t_2)^2\rangle - \langle Y(t_2)\rangle^2}}$$

Pour les corrélations croisées « vert-rouge », on notera tV et tR les temps auxquels sont évaluées les intensités de fluorescence verte et rouge respectivement. Les intensités aux deux temps sont évaluées dans une même lignée : si par exemple t_1 est supérieure à t_2, pour chaque bactérie considérée à t_1 l'intensité à t_2 est celle de son ancêtre à cet instant.

Ces corrélations donnent les dépendances linéaires de deux variables : c'est une première indication de leurs régulations mutuelles. On remarquera que par définition les autocorrélations sont symétriques et valent 1 sur la première bissectrice dans le plan (t_1, t_2). Pour les

7. Il serait aussi intéressant d'avoir une estimation de leur taux de production, mais les dérivées des intensités de fluorescence mesurées ici sont trop bruitées pour qu'on puisse en tirer des informations.

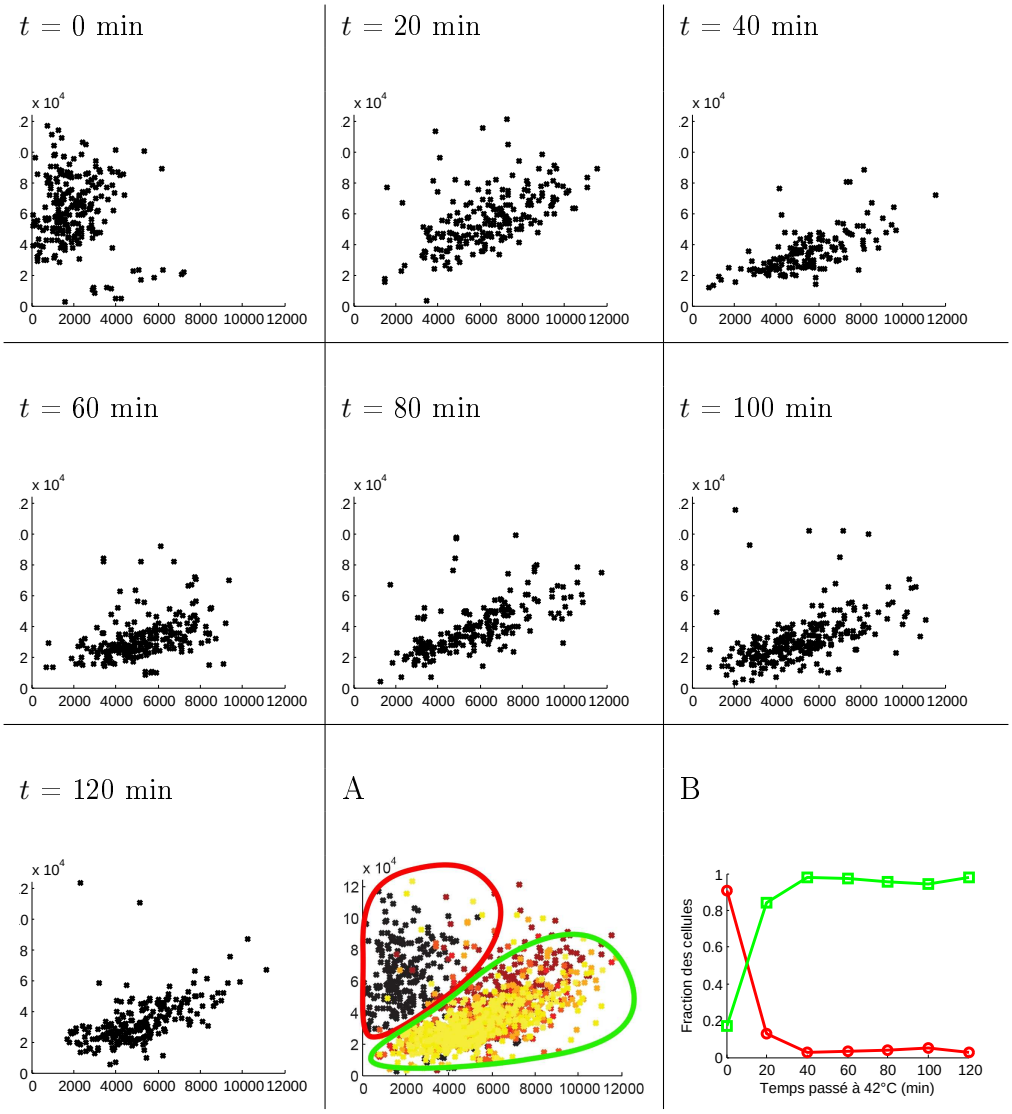


Fig. 2.7 – Légende identique à la légende de la figure 2.5, pour des bactéries ayant passé un temps t à 42 ° C dans du PBS, puis ont poussé 2h dans du LB à 30 ° C. On remarque une fraction de bactéries « lytiques » à $t = 0$ min.

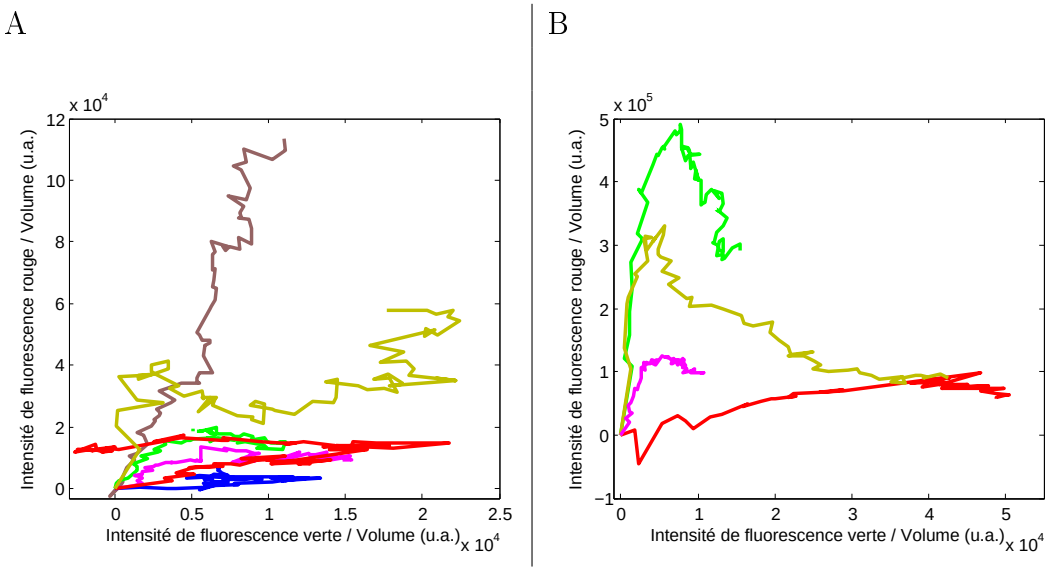


Fig. 2.8 – Diagramme d'évolution temporelle des moyennes des intensités de fluorescence divisées par le volume de bactéries (A) pZC-Lf1 et (B) pMK-Lf1. En (A) : courbe brune : sans choc de température, sur du MM ; courbe bleue : sans choc de température, sur du LB à 37 ° C. En (A) et (B) : courbes rouges : OD 2,5 sur du LB ; courbes vertes : OD 2,5 sur du MM ; courbes jaunes : OD 0,05 sur du LB ; courbes magenta : OD 0,05 sur du MM.

corrélations croisées, les valeurs en-dessous de la première bissectrice dans le plan (tV, tR) peuvent être interprétées comme mesurant une « action du rouge sur le vert » (tR est plus petit que tV), c'est-à-dire ici une action de CI sur la production de Cro ; et inversement pour les valeurs au-dessus de la première bissectrice.

Dans le paragraphe suivant, les résultats de mesures sans choc de température sont présentées : dans des conditions favorables à la maintenance de la lysogénie (à 30 ° C sur du MM) et dans des conditions forçant le basculement de la lysogénie vers la lyse (à 37 ° C sur du LB). Le paragraphe 2.4.2 expose les mesures d'expression après dénaturation des répresseurs dans huit conditions : deux de densité, deux de milieu de croissance et deux de nombre de copies du réseau.

Les diagrammes des moyennes des intensités de fluorescence corrigées et divisées par le volume, dans ces différentes conditions, sont indiqués sur la figure 2.8.

Remarque : j'appellerai « famille » de bactéries l'ensemble des descendantes d'une même bactérie au début de la mesure : à conditions données, il y aura autant de familles que de bactéries à l'instant initial.

2.4.1 Sans dénaturation préalable des répresseurs CI

Des bactéries lysogènes (colonie apparaissant rouge sur boîte inoculée dans du MM et incubée une nuit à 30 ° C) ont été étalées sur de l'agar supplémenté en MM et leur croissance et fluorescence mesurées pendant 20h. La figure 2.9 présente ces mesures pour les six premières heures d'observation. Le champ observé contenait initialement six bactéries ; deux d'entre elles apparaissant vertes et pas rouges, et inversement pour les quatre autres. Au cours des vingt heures de film, toutes les descendantes de ces deux bactéries ont gardé des niveaux de fluorescences verte élevée et rouge faible, et inversement pour les autres. Cela se voit bien sur les figures 2.9 A et C.

Les corrélations peuvent être interprétées dans le même sens : les autocorrélations (figure 2.9 B et D) prennent des valeurs proches de 1 presque partout : chaque « état » semble stable ; de plus, les corrélations croisées (figure 2.9 E) prennent des valeurs proches de -1 presque partout : les deux états s'excluent mutuellement.

On a ainsi une nouvelle indication de la fidélité du marquage par les gènes codant pour des protéines fluorescentes ; les deux comportements observées s'interprètent bien comme lysogénique et lytique. Cet exemple semble de plus indiquer que la lyse est très stable, même dans des conditions favorables à la lysogénie ; il faudrait cependant davantage d'observations de ce type pour le confirmer.

Remarque : les moyennes des intensités ont été indiquées, bien qu'elles reflètent mal ce comportement bimodal.

Remarque : il ne s'agit pas de bactéries qu'on verrait « choisir » l'une ou l'autre voie ; ici, la correction des données masque quelque peu le fait qu'un état stationnaire est vraisemblablement atteint avant le début de l'observation : il s'est avéré que la proportion de lysogènes n'était pas de un ; on ne peut pas dire si ces bactéries lytiques sont les descendantes de bactéries lytiques ou si une induction de lyse a eu lieu.

Des bactéries préparées de la même manière mais mises à pousser sous le microscope à 37 ° C sur de l'agar supplémenté en LB ont ensuite été observées ; on devrait ainsi voir le passage (forcé) de la lysogénie vers la lyse, c'est-à-dire le transitoire entre les deux états décrits à la figure 2.4. Ces mesures sont présentées sur la figure 2.10. Le niveau de fluorescence rouge (figure 2.10 C) reste faible comparé aux niveaux hauts de l'expérience précédente. La fluorescence verte (2.10 A) croît initialement rapidement, plus vite que pour les bactéries lytiques de l'expérience précédente ; cependant, à part pour une famille (en rouge) et certaines bactéries d'une autre (cyan), qui se maintiennent à un niveau élevé, la plupart des bactéries voient leur fluorescence verte s'effondrer après une centaine de minutes. On peut s'attendre à ce qu'un niveau de fluorescence verte plus haut que le niveau de l'état stationnaire en régime transitoire, Cro réprimant sa propre production à haute concentration ; cela explique cependant mal que la décroissance soit si abrupte. On peut penser aussi à un changement physiologique des cellules : la plupart, au cœur de la

colonie, croissant de moins en moins vite et ayant des tailles assez réduites.

Les corrélations semblent structurées en deux zones : un carré $t_1, t_2 < 130$ min et le reste ; cela vient de ce que les données ont été obtenues à partir de trois expériences, dont l'une n'a duré que 130 min. On remarque ainsi que des variations importantes peuvent exister, dans les mêmes conditions, d'une expérience à l'autre et qui n'apparaissent pas clairement sur les courbes d'intensité. L'autocorrélation verte (2.10 B) prend des valeurs proches de 1 sur les deux carrés $t_1, t_2 < 130$ min et $t_1, t_2 > 130$ min, encore une fois cohérente avec un comportement (bi)stable ; les valeurs en dehors de ces deux carrés sont plus difficiles à interpréter. En dehors des pics aux coins bas-droite et haut-gauche, que je n'explique pas, la figure d'autocorrélation rouge est cohérente avec un répresseur désactivé : elle tombe rapidement de 1 à 0 quand on s'éloigne de la bissectrice (2.10 D).

Les corrélations croisées (2.10 E) portent peut-être la trace d'une répression de l'expression de *cI* par Cro en $(tV, tR) \simeq (50, 75)$, alors que CI ne peut réprimer l'expression de *cro* ($(tV, tR) \simeq (100, 50)$).

2.4.2 La décision lyse/lysogénie

Le protocole suivant permet de simuler l'infection d'une bactérie par Lambda et de suivre la décision entre lyse et lysogénie dans huit conditions (deux de nombre de copies, deux de densité de bactéries, deux de milieu de croissance) :
- inoculer dans du MM une colonie de bactéries pZC-Lf1 ou pMK-Lf1 apparaissant rouge [8] ;
- les incuber une nuit à 30 ° C ;
- les diluer 2 ou 100 fois dans du MM ;
- les incuber 3h à 30 ° C, de telle sorte que les culture atteignent des densités optiques à 600 nm (« OD ») d'environ 2,5 ou 0,05 respectivement ;
- les incuber 30 min au bain-marie à 42 ° C ;
- les étaler sur de l'agar supplémenté en LB (milieu riche) ou MM (milieu pauvre) ;
- mesurer leur croissance et fluorescences verte et rouge au cours du temps, l'échantillon étant maintenu à 30 ° C.

Les résultats sont présentés sur les figures 2.11 à 2.18, organisées par nombre de copies, densité puis milieu. Ces données sont plus difficiles à interpréter que celles présentées au paragraphe précédent. Aucune image nette ne semble s'en dégager ; je commenterai séparément les résultats obtenus dans ces différentes conditions.

Pour le plasmide à copie unique, à forte densité, le comportement observé sur les courbes d'intensités dépend peu du milieu de croissance (figures 2.11 et 2.12), si ce n'est que les variations dans les niveaux d'expression de *cro* semblent plus importantes sur du LB que sur du MM. Il ressemble à ce qu'on a vu pour des bactéries pZC-Lf1 croissant à 37 ° C sur du

8. Rappelons que : 1. le plasmide pZC a un nombre de copies moyen de 1, le plasmide pMK de plusieurs dizaines ; 2. une culture dans du MM à 30 ° C permet le maintien de la lysogénie.

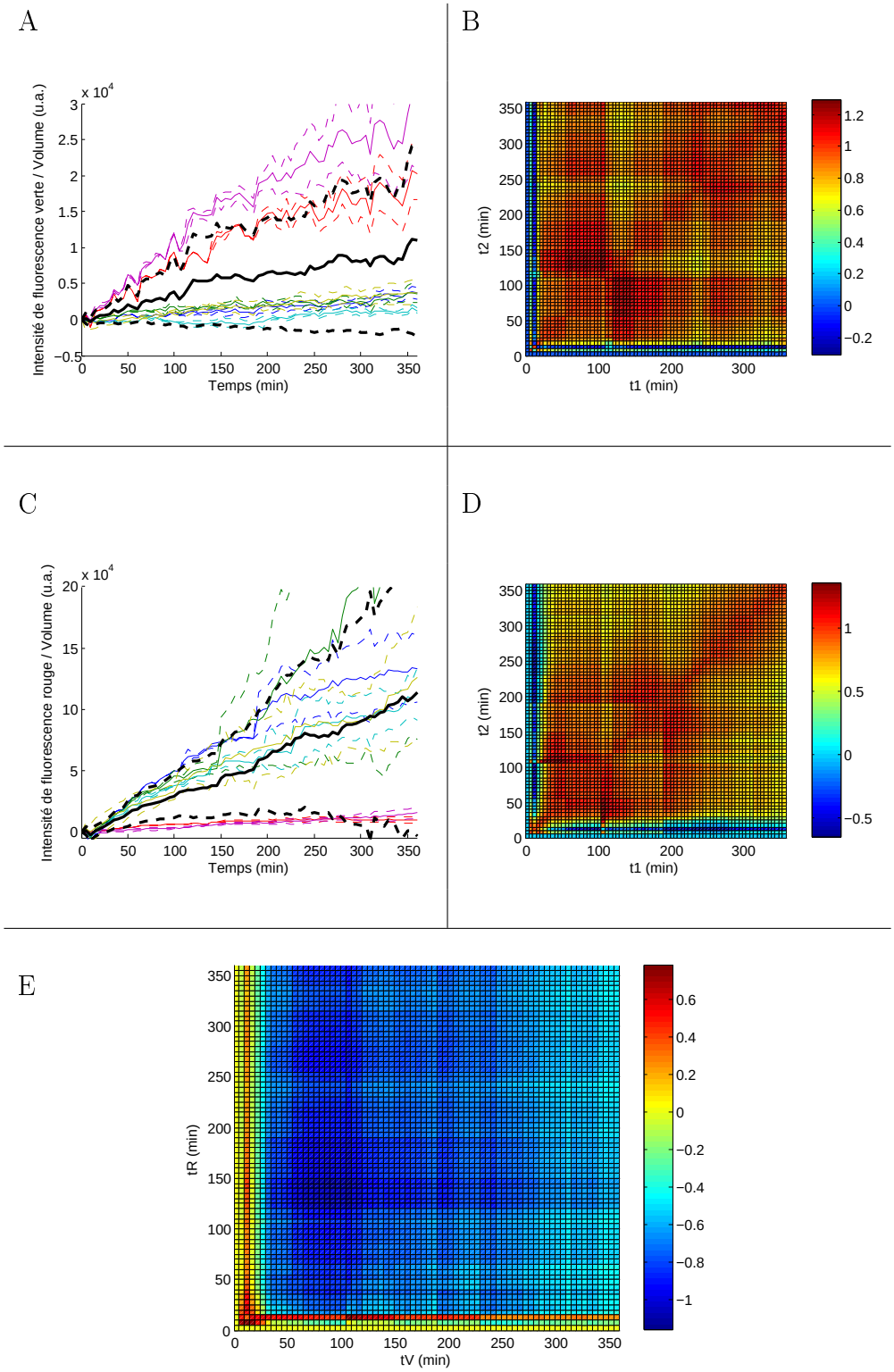


Fig. 2.9 – Mesures de fluorescence de pZC-Lf1, sur du MM à 30 ° C, sans choc de température. (A) Intensité de fluorescence verte divisée par le volume, corrigée comme indiqué en annexe A.4.2, en fonction du temps ; en couleur : familles distinctes de bactéries ; en noir : ensemble des bactéries ; courbes pleines : moyennes ; courbes tiretées : moyennes plus ou moins une déviation standard. (B) autocorrélation à différents temps de l'intensité verte divisée par le volume corrigée ; (C) et (D) : comme (A) et (B), pour la fluorescence rouge. La même couleur est attribuée à la même famille sur les figures (A) et (C). (E) Corrélation croisée vert-rouge à différents temps. Les corrélations sont définies dans le corps du texte. *Remarque* : sur les figures (B), (D) et (E), les spectres de couleurs correspondent à différentes gammes de valeurs.

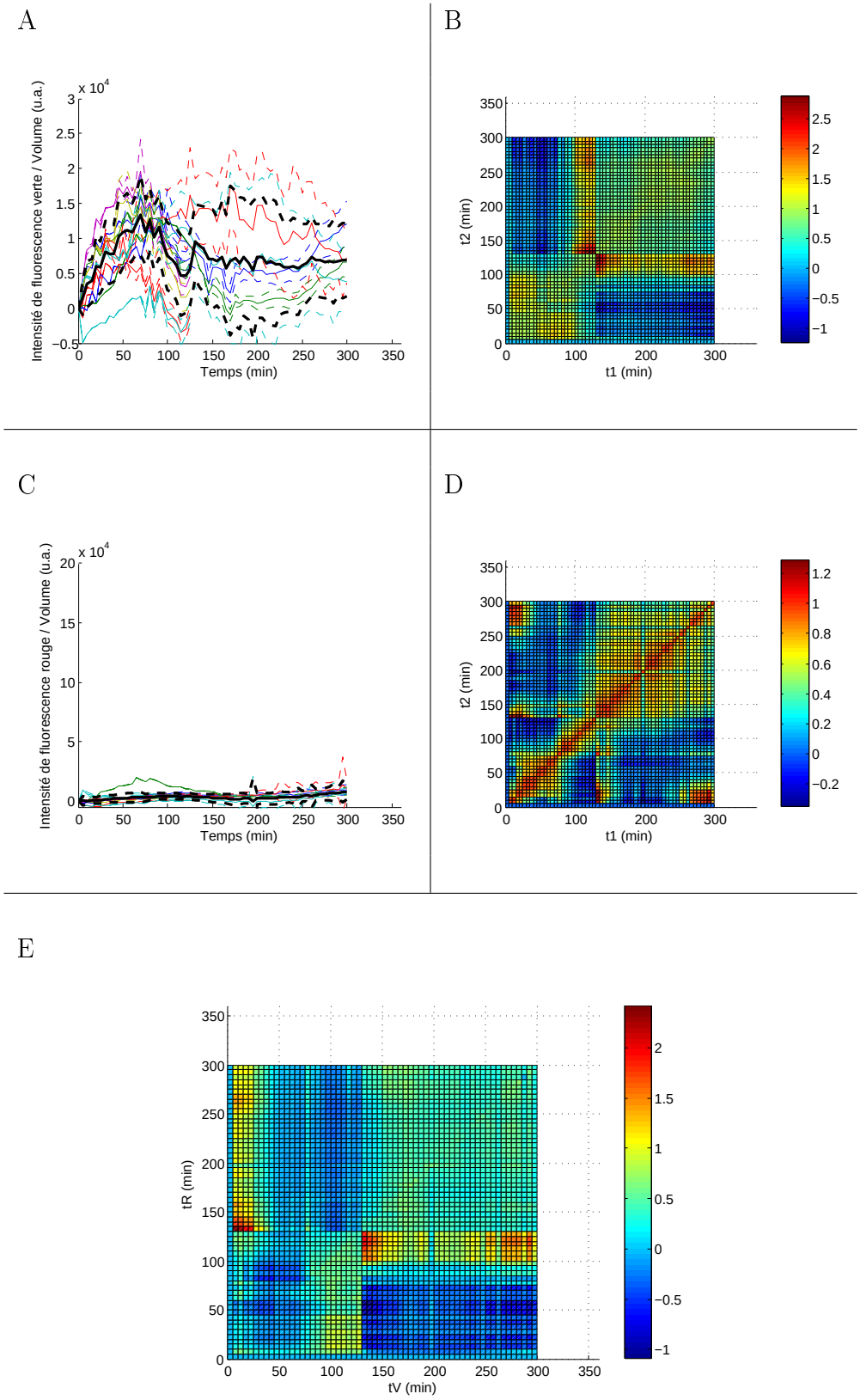


Fig. 2.10 – Mesures de fluorescence de pZC-Lf1, sur du LB à 37 ° C, sans choc de température. Voir la figure 2.9 pour le détail de la légende.

LB : la fluorescence verte croît pendant un temps (ici les 200 premières minutes environ), avant de diminuer, brutalement dans le cas d'un milieu riche ; la fluorescence rouge elle reste faible, sauf pour une bactérie dans le cas d'un milieu pauvre qui la voit croître pendant les 100 premières minutes avant de revenir à un niveau faible. Ainsi, la plupart des bactéries semblent suivre une voie lytique, avec peut-être une tentative avortée d'établissement de lysogénie ; cela est cohérent avec le comportement attendu, très majoritairement une lyse sur milieu riche, la possibilité de suivre la lysogénie sur un milieu pauvre.

Des corrélations présentées sur ces deux figures, seules les corrélations croisées dans le cas d'un milieu pauvre présentent une structure claire (figure 2.12 E) : nulles le plus souvent, négatives pour tV inférieur à 100 min et tR proche de 150 min, proches de 1 pour tR inférieur à 50 min et tV inférieur à 150 min ; cela suggérerait, au début de l'observation, une répression de *cI* par Cro et une activation de *cro* par CI. Cette activation ne correspond cependant pas au comportement attendu, une répression de *cro* par CI.

Pour une faible densité et des bactéries croissant sur de l'agar supplémenté en LB, pour les deux plasmides (figures 2.13 et 2.17), la fluorescence verte croît sur l'ensemble de la durée d'observation alors qu'on voit un maximum de fluorescence rouge atteint au bout de 50 à 100 minutes. Dans le cas d'un faible nombre de copies cependant, un comportement bimodal semble émerger après 150 minutes (sur les courbes d'intensité 2.13 A et C, et surtout dans le carré supérieur droit de 2.13 E, où les corrélations croisées prennent des valeurs proches de -1), qu'on ne constate pas dans le cas d'un haut nombre de copies.

De façon étonnante, dans le cas d'un haut nombre de copies, d'une densité forte et d'un milieu de croissance pauvre (figure 2.16), alors qu'on distingue bien trois comportements en fluorescence verte (courbes d'intensité et autocorrélation), on n'en voit pas la trace en fluorescence rouge ni sur les corrélations croisées. Quelques bactéries semblent suivre une voie lytique avant de « se raviser ». Un temps caractéristique de 150 minutes semble ici encore se dessiner.

Enfin, dans les conditions de faible densité, sur du MM, pour les deux vecteurs (figures 2.14 et 2.18) et pour le plasmide pMK à haute densité sur du LB (figure 2.15), le faible nombre de bactéries observées ne permet pas de proposer une interprétation.

2.4.3 Conclusion

Les mesures réalisées sans choc préalable de température confirment que les constructions (réseau de régulation de Lambda isolé, fusion transcriptionnelle des gènes *egfp* et *mCherry* derrière *cro* et *cI*) permettent de distinguer des comportements « lytique » et « lysogénique », et d'étudier ces deux régimes permanents ; en particulier, la lyse semble stable sur un vingtaine d'heures, ou une vingtaine de générations, dans des conditions défavorables.

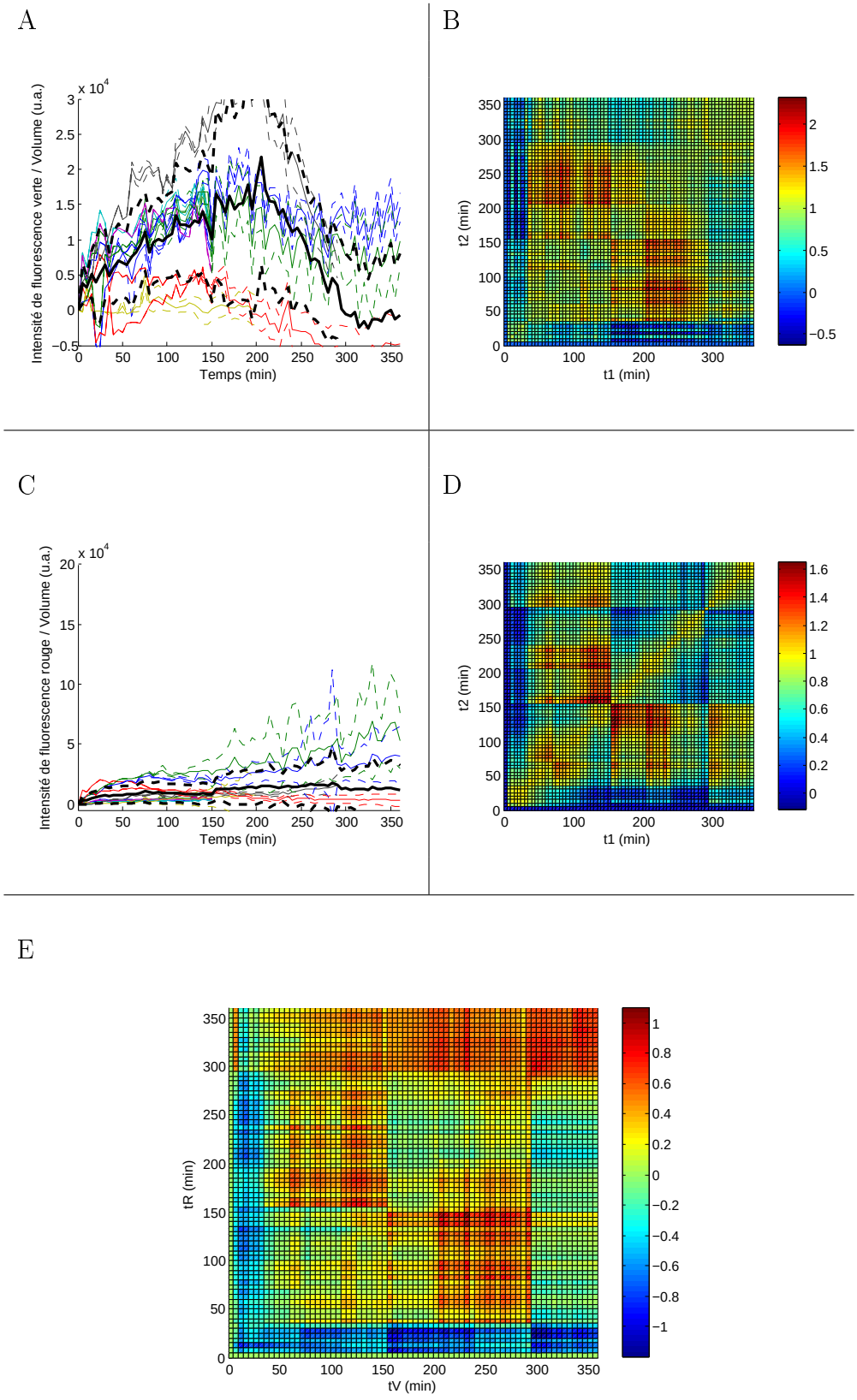


Fig. 2.11 – Mesures de fluorescence, après choc de température, de **pZC-Lf1**, à **OD 2,5** environ, sur du **LB**. Voir la figure 2.9 pour le détail de la légende.

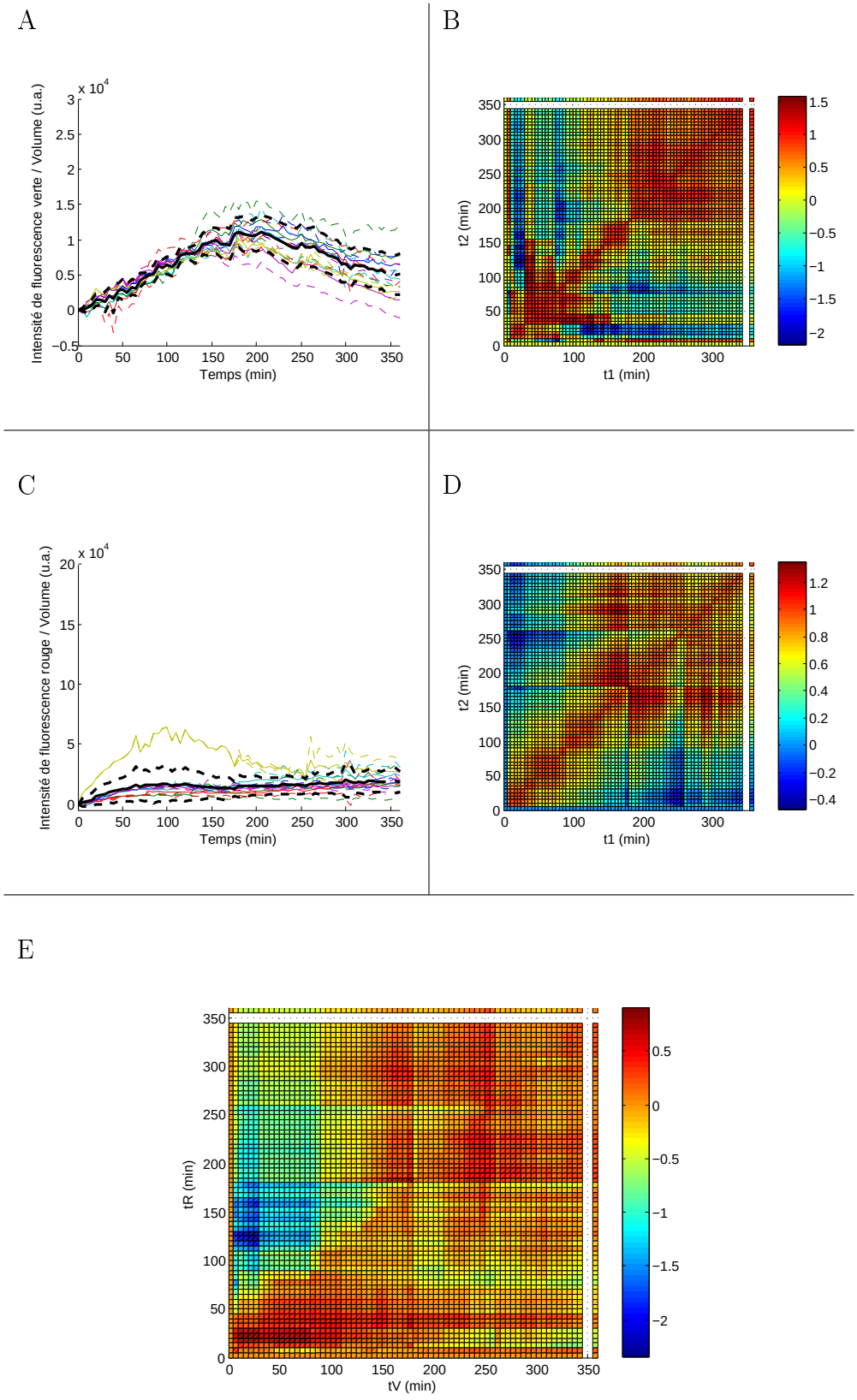


Fig. 2.12 – Mesures de fluorescence, après choc de température, de **pZC-Lf1**, à **OD 2,5** environ, sur du **MM**. Voir la figure 2.9 pour le détail de la légende.

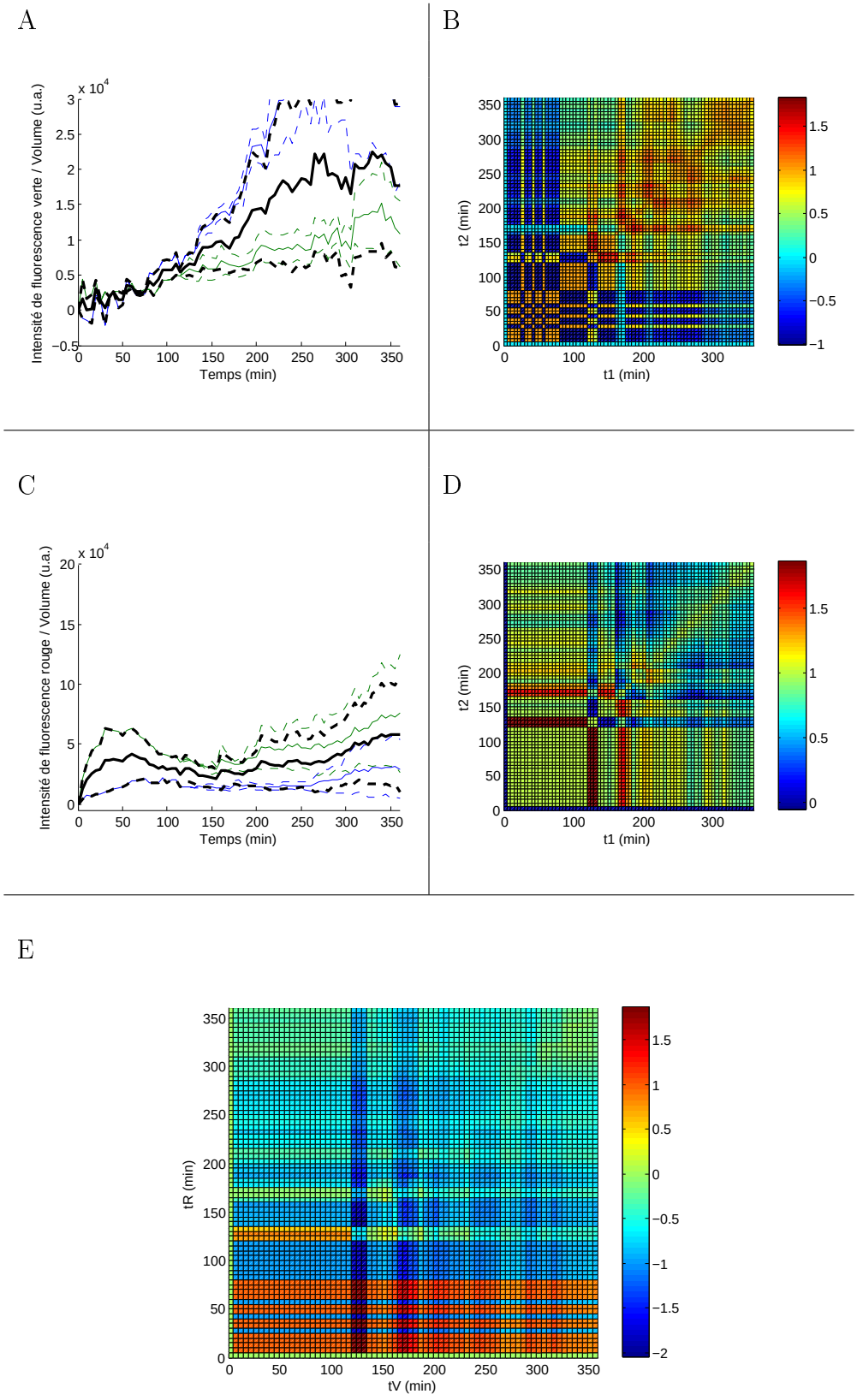


Fig. 2.13 – Mesures de fluorescence, après choc de température, de **pZC-Lf1**, à **OD 0,05** environ, sur du **LB**. Voir la figure 2.9 pour le détail de la légende.

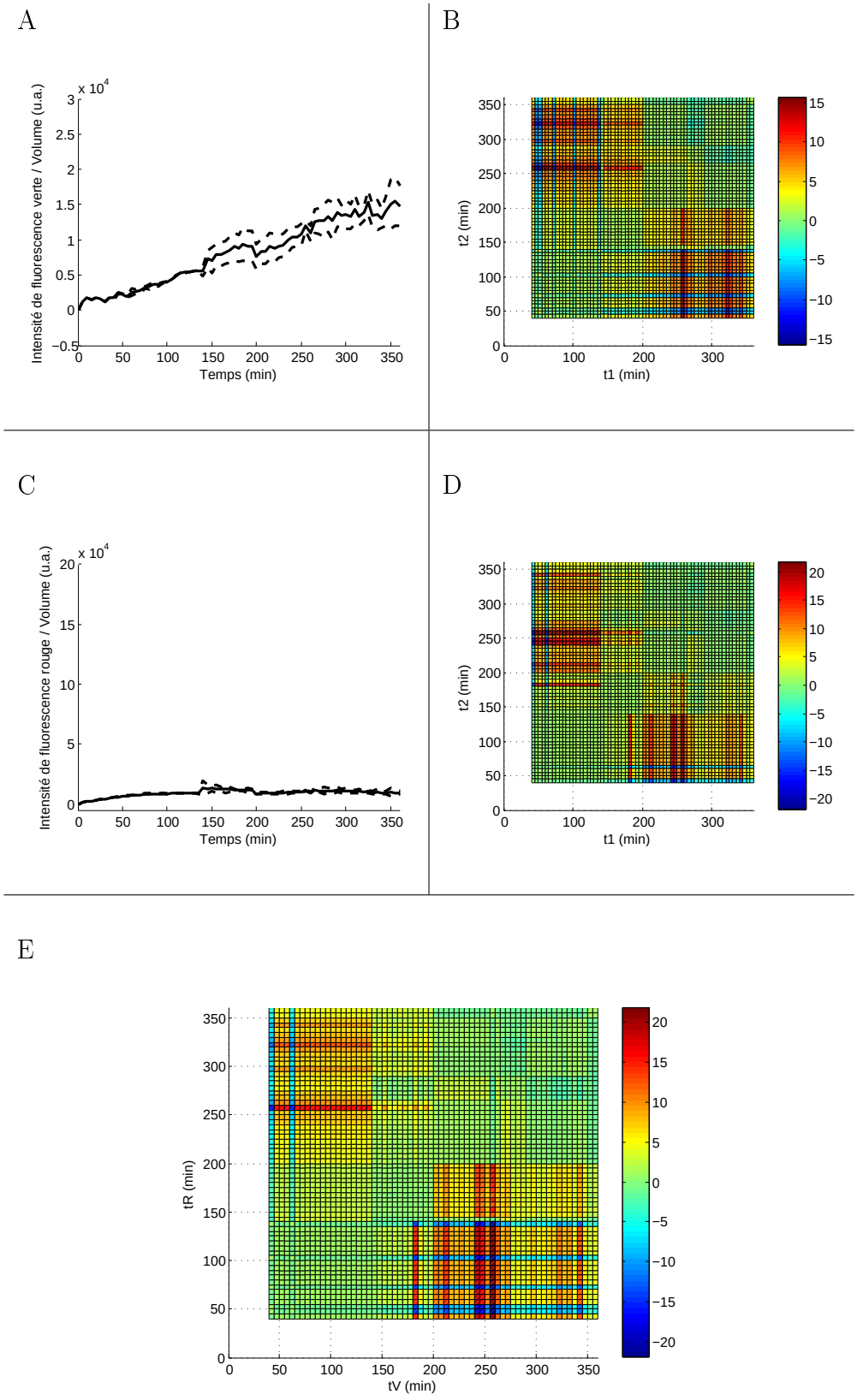


Fig. 2.14 – Mesures de fluorescence, après choc de température, de **pZC-Lf1**, à **OD 0,05** environ, sur du **MM**. Voir la figure 2.9 pour le détail de la légende.

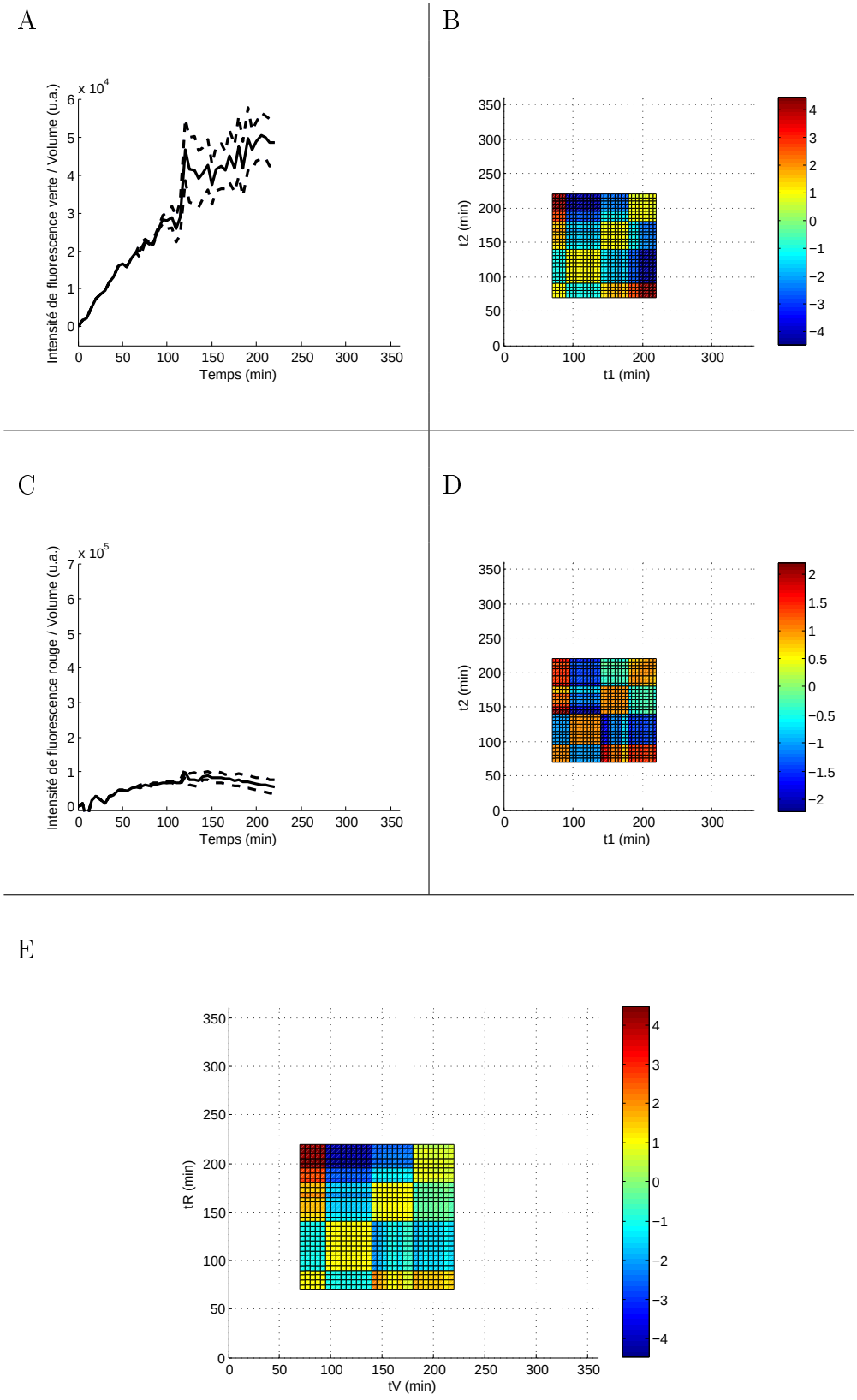


Fig. 2.15 – Mesures de fluorescence, après choc de température, de **pMK-Lf1**, à **OD 2,5** environ, sur du **LB**. Voir la figure 2.9 pour le détail de la légende.

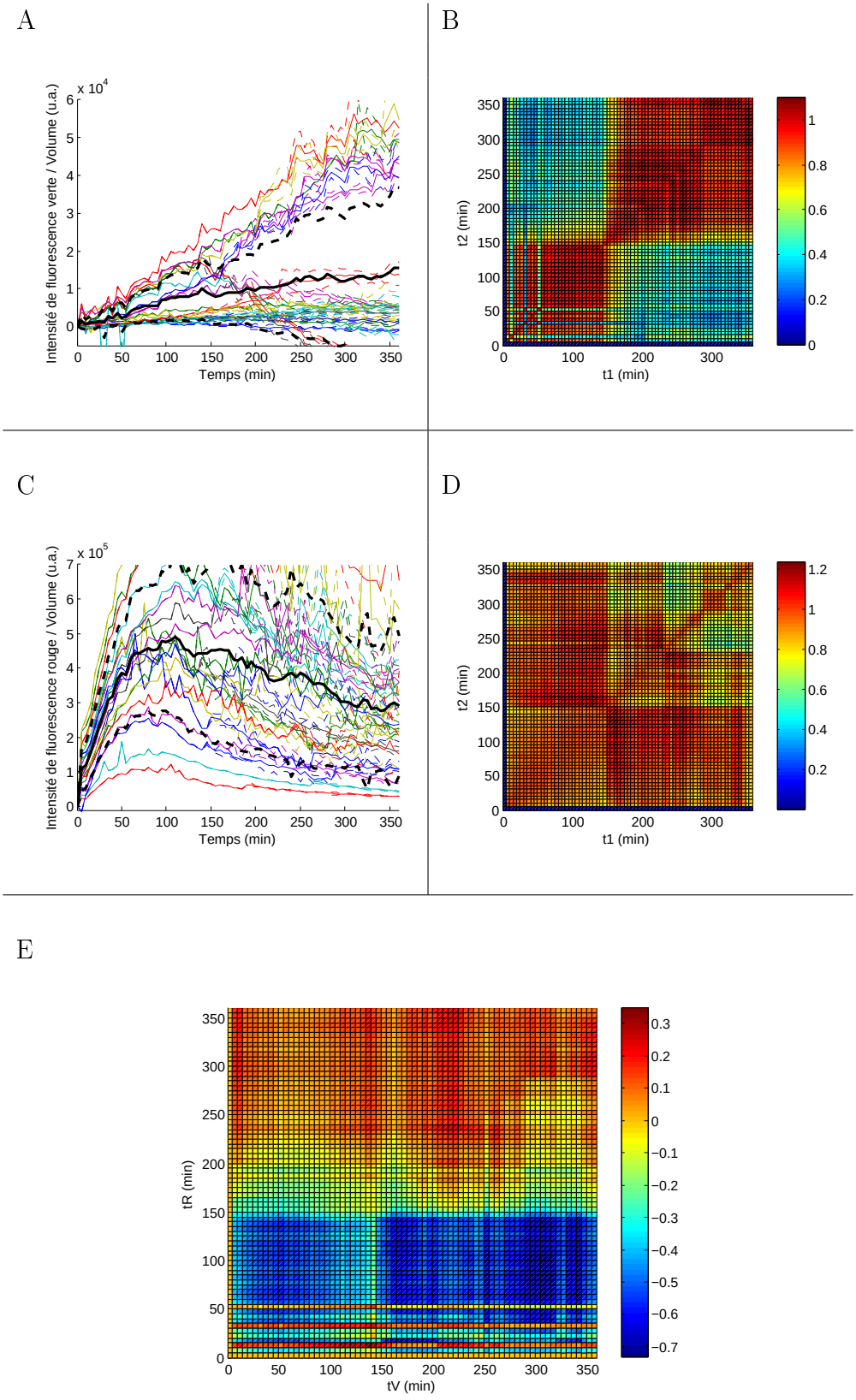


Fig. 2.16 – Mesures de fluorescence, après choc de température, de **pMK-Lf1**, à **OD 2,5** environ, sur du **MM**. Voir la figure 2.9 pour le détail de la légende.

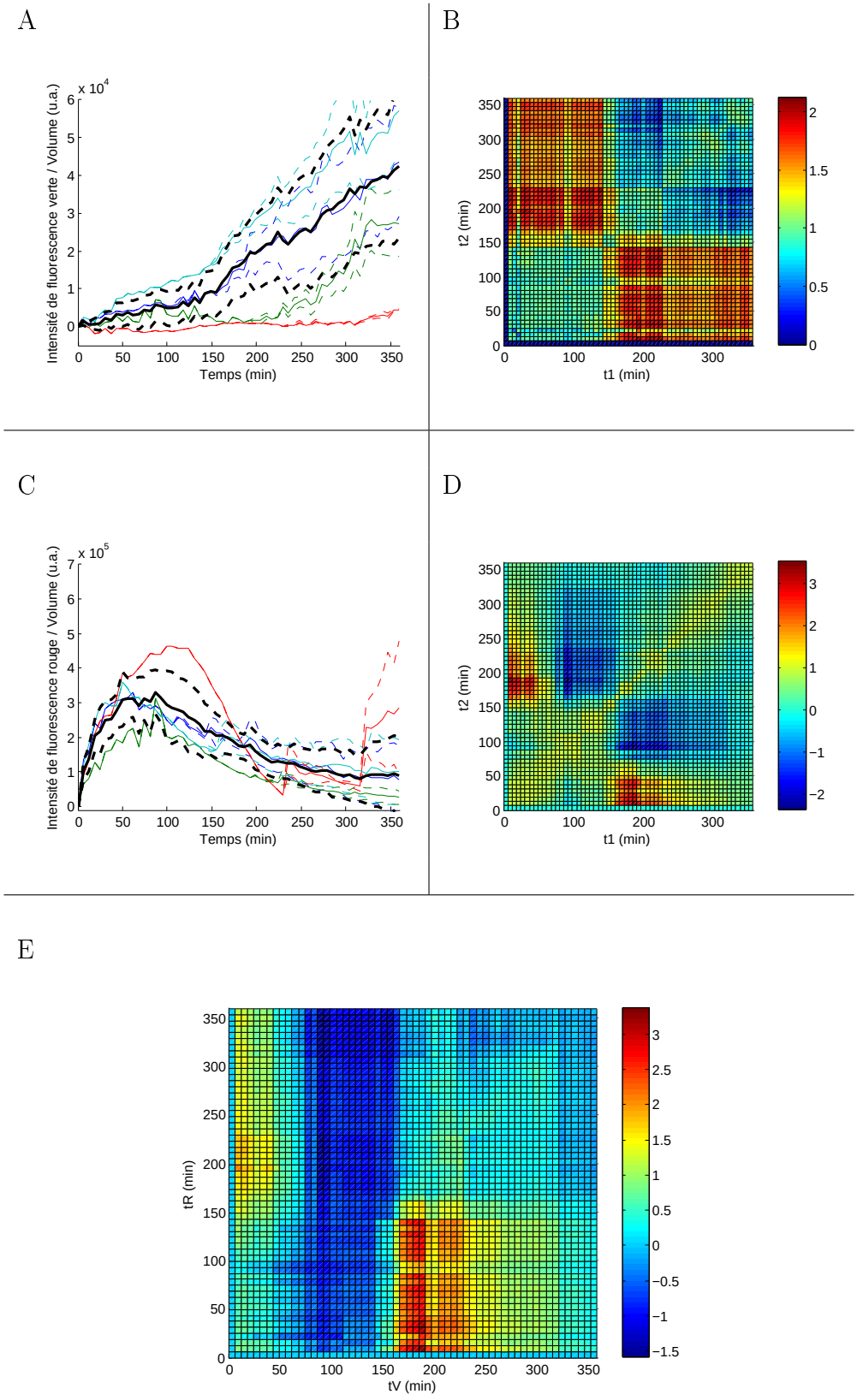


Fig. 2.17 – Mesures de fluorescence, après choc de température, de **pMK-Lf1**, à **OD 0,05** environ, sur du **LB**. Voir la figure 2.9 pour le détail de la légende.

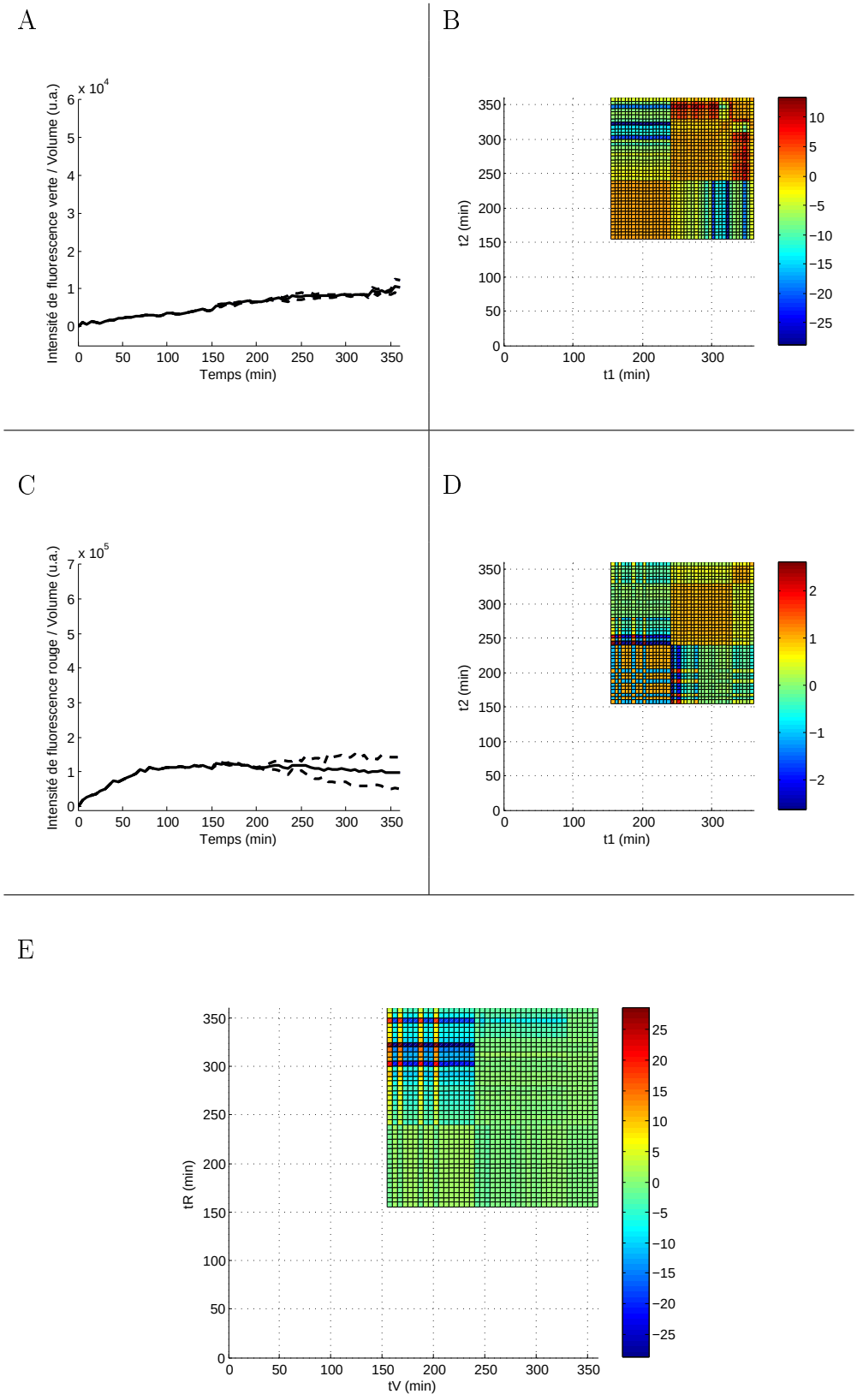


Fig. 2.18 – Mesures de fluorescence, après choc de température, de **pMK-Lf1**, à **OD 0,05** environ, sur du **MM**. Voir la figure 2.9 pour le détail de la légende.

Par contre, avec les données dont nous disposons pour l'instant, l'étude de la décision est plus difficile. Le nombre d'expériences est souvent insuffisant pour distinguer un comportement significatif, d'autant plus que la forte perturbation du passage de trente minutes à 42 ° C peut masquer, sur un faible nombre d'échantillons, le biais systématique provenant des conditions de croissance ou du nombre de copies du réseau. Il serait intéressant de voir si la durée de ce choc de température pourrait être réduite. Cette levée de répression pourrait être efficacement contrôlée, avec un impact minimal sur les cellules, en utilisant un dispositif microfluidique de changement de milieu en cours d'observation [49], en faisant passer des milieux de températures différentes.

Un plus grand nombre de mesures permettrait de plus de s'intéresser aux taux de production, et d'étudier la corrélation entre taille ou taux de croissance des bactéries et comportement du réseau.

Le dispositif expérimental pourrait aussi être amélioré de façon à ce que les bactéries poussent mieux au cours de l'observation. En effet, j'ai constaté que les cellules au centre des colonies croissent moins vite et sont plus petites que celles au bord. L'utilisation de canaux creusés dans l'agar ou d'un dispositif microfluidique assurant un apport continu de nutriments pourraient corriger ce problème.

Des développements immédiats consisteraient à mesurer les fractions de bactéries lytiques et lysogéniques au sein d'une population, en fonction des conditions de croissance[9], et, bien sûr, à répéter les mesures présentées avec les constructions Lf2 et Lf3.

Il serait aussi intéressant de faire ces mesures avec des fonds génétiques de la bactérie hôte différents, ou, mieux, de placer des facteurs d'hôte (RNaseIII, HflB, RecA) sous promoteur inductible et de mesurer la « réponse » de Lambda, soit en présence de différentes concentrations d'inducteur, soit à un apport bref d'inducteur.

9. Une telle mesure quantitative fait grandement défaut ; la seule étude de référence est [50].

Chapitre 3

Conception de réseaux artificiels

Le réseau de régulation de Lambda remplit une fonction simple, qu'on peut qualifier de commututateur biaisé. Il apparaît à la fois complexe et « optimisé » (compact sur le génome, bien conservé entre bactériophages lambdoïdes). Comparer les mesures de dynamique d'expression de ce réseau au comportement de réseaux artificiels, générés sur ordinateur, qui seraient capables de remplir une même fonction pourra aider à mieux comprendre sa forme particulière et peut-être mettre en lumière des caractéristiques jusque-là ignorées.

À la limite, on peut imaginer des expériences où le réseau de décision de Lambda est remplacé par les réseaux ainsi conçus et étudier à quel point un tel remplacement aboutit à un virus fonctionnel[1].

Genherite est un programme écrit par Paul François et Vincent Hakim qui permet de générer sur ordinateur des réseaux de régulation génétique et de simuler leur dynamique [52, 53]. Je présente dans la suite succinctemment son principe de fonctionnement puis les composants que j'y ai ajoutés, de manière à inclure les types de régulation rencontrés chez Lambda. J'ai travaillé sur une version du programme réécrite par Hervé Rouault.

3.1 Présentation succincte de *Genherite*

3.1.1 Modélisation des réseaux génétiques

Les réseaux génétiques sont définis ici par un ensemble de réactions chimiques élémentaires, ayant lieu en même temps, entre des gènes, les ARN correspondants et des protéines[2]. Elles sont modélisées par des équations différentielles du premier ordre, qui rendent bien compte de l'évolution temporelle du nombre moyen de molécules.

1. Dans l'esprit de [51], où les auteurs ont remplacé cI par $tetR$, cro par $lacI$, O_{R3} par $lacO$ et O_{R1} et O_{L1} par $tetO$ et $lacO$, et ont obtenu des phages « viables ».

2. Les processus moléculaires sont en réalités beaucoup plus nombreux, cette représentation néglige notamment les étapes de polymérisation, leur initiation, les interactions avec le solvent ou les ions, etc. Ils sont cependant beaucoup plus rapides que les variations typiques de concentration de protéines ou d'ARN.

Les réactions considérées sont les :
- transcription (gène $a \rightarrow$ gène a + ARN a) ;
- traduction (ARN a $\rightarrow$ ARN a + protéine A) ;
- dégradation (ARN a $\rightarrow \emptyset$ ou protéine A $\rightarrow \emptyset$) ;
- phosphorylation (protéine A $\rightarrow$ protéine phosphorylée A*) ;
- dimérisation (protéine A + protéine B $\rightarrow$ protéine AB) ;
- dissociation (protéine AB $\rightarrow$ protéine A + protéine B) ;
- dégradation active par une protéine du réseau (protéine A + protéine B $\rightarrow$ protéine A).

De plus, les transcriptions peuvent être régulées par des protéines se fixant à l'ADN. Les réactions d'association et dissociation d'une protéine sur l'ADN sont supposées à l'équilibre : elles peuvent être intégrées dans un taux de transcription effectif qui dépend du nombre de protéines régulatrices[3].

Le nombre de copies d'un gène a reste constant. Il est pris égal à 1.

Voici par exemple les équations décrivant la dynamique d'un réseau constitué d'un seul gène a, transcrit en ARN a et dont le produit A forme un dimère AA qui en régule la transcription et peut se dissocier ($[X]$ désigne le nombre de molécules d'espèce X, éventuelle-

3. Considérons un promoteur p, sous lequel est transcrit un ARN a et auquel peuvent s'attacher des protéines P_i. Soit pP_i le complexe formé par le promoteur et la protéine P_i, k_l et $k_{o,i}$ les taux de transcription lorsque le promoteur est libre et occupé par la protéine P_i, respectivement. Soit t_i^+ et t_i^- les constantes d'association et dissociation de la protéine P_i sur le promoteur. En considérant qu'une seule protéine peut se fixer à la fois au promoteur, on peut écrire ($[X]$ désigne le nombre de molécules d'espèce X, éventuellement inférieur à 1) :

$$\begin{aligned} \frac{d[a]}{dt} &= k_l[p] + \sum_i k_{o,i}[pP_i] \\ \frac{d[pP_i]}{dt} &= t_i^+[p][P_i] - t_i^-[pP_i] \end{aligned}$$

Si l'on suppose les réactions sur le promoteur à l'équilibre et $[p]+\sum_i[pP_i] = 1$ (une seule molécule d'ADN), alors :

$$[pP_i] = K_i[p][P_i] \text{ et } [p] = \frac{1}{1+\sum_i K_i[P_i]} ,$$

où $K_i := t_i^+/t_i^-$ est la constante d'équilibre de la réaction d'association de la protéine P_i au promoteur. Il s'ensuit :

$$\frac{d[a]}{dt} = \frac{k_l + \sum_i k_{o,i} K_i[P_i]}{1+\sum_i K_i[P_i]}$$

Remarque : on trouve ainsi que, si une seule copie de chaque protéine régule la transcription, le taux de transcription est une généralistation d'une fonction de Hill de degré 1.

Il est en principe possible de prendre en compte le fait que des protéines de types différents peuvent se fixer en même temps sur un même promoteur, mais les nombres de cas à envisager, de constantes et d'équations croissent alors énormément. Le cas de plusieurs protéines du même type est plus simple (voir la section 3.2, le paragraphe « Coopérativité »).

ment inférieur à 1) :

$$\begin{aligned}
\frac{d[a]}{dt} &= \frac{k_l + k_o K[AA]}{1 + K[AA]} - \delta_a[a] \\
\frac{d[A]}{dt} &= k_A[a] - k_A^+[A]^2 + k_{AA}^-[AA] - \delta_A[A] \\
\frac{d[AA]}{dt} &= k_A^+[A]^2 - k_{AA}^-[AA] - \delta_{AA}[AA]
\end{aligned}$$

où k_l est le taux de transcription « libre » (c'est-à-dire quand AA n'est pas fixé à l'ADN), k_o le taux de transcription « occupé »(AA fixé à l'ADN), K la constante d'équilibre de la réaction d'association de AA à l'ADN, δ_X le taux de dégradation de l'espèce X, k_A le taux de traduction, k_A^+ le taux de dimérisation et k_{AA}^- le taux de dissociation.

Remarque : si k_o est plus grand que k_l, on dit que AA active (ou promeut) la transcription du gène a, dans le cas contraire qu'elle la réprime (ou inhibe).

3.1.2 L'algorithme d'évolution (mutation et sélection)

Le programme fait « évoluer » N_r réseaux identiques, définis par l'utilisateur, selon l'algorithme d'évolution suivant (chaque étape est expliquée dans la suite) :
1. appliquer deux « mutations » (tirées aléatoirement) à chacun de la deuxième moitié d'entre eux ;
2. simuler le comportement de tous les réseaux (intégrer les équations couplées définissant chaque réseau) ;
3. leur attribuer un score, selon une « fonction de score » prédéfinie ; les trier par ordre de score décroissant ;
4. supprimer la deuxième moitié (celle de score le moins bon), dupliquer la première ;
5. revenir en 1.

L'algorithme est appliqué N_g fois. Une procédure d'optimisation (voir ci-dessous) est appliquée au meilleur réseau ainsi obtenu.

Les « mutations » possibles consistent à :
- modifier les quantités initiales d'ARN ou de protéines ;
- ajouter ou supprimer un gène ;
- ajouter ou supprimer une réaction ;
- ajouter ou supprimer la régulation d'un gène par une protéine ;
- modifier une constante cinétique (de réaction, de dégradation, de transcription « libre », « occupé », d'équilibre d'association à l'ADN).

La probabilité d'appliquer un type de mutation est définie par l'utilisateur ; en pratique il est plus efficace de modifier plus souvent les constantes cinétiques que la topologie du

réseau.

Le score sert à distinguer les réseaux s'approchant du comportement souhaité par l'utilisateur. Il n'y a pas de règle pour son calcul. Dans la mesure où c'est la dynamique des réseaux générés qui nous intéresse, on prendra naturellement comme score l'opposé de la distance entre une fonction cible (comportement désiré) et le nombre d'une protéine au cours du temps, mais des fonctions plus subtiles peuvent s'avérer plus efficaces (par exemple, considérer les corrélations temporelles du nombre d'une protéine).

Pour éviter de générer des réseaux inutilement complexes, une pénalité est appliquée : plus un réseau aura de gènes (et donc d'ARN) et de protéines, moins, à comportement donné, son score sera bon. En pratique, le score est multiplié par un multiple du nombre d'espèces (quand il est attribué de telle façon qu'il est négatif).

La prodédure d'optimisation est identique à l'algorithme décrit plus haut, à la différence qu'elle n'est pas répétée le même nombre de fois, que les « mutations » complexifiant le réseau (ajout de gène, régulation, réaction) ne sont pas considérées et que les probabilités de tirer une mutation ne sont pas les mêmes.

La figure 3.1 présente un exemple de fonctionnement de *Genherite*.

3.2 Les composants ajoutés

3.2.1 Clivage de protéines et dégradation active d'ARN

Lorsqu'il existe des complexes, une réaction de clivage peut être créée : AB + C → A + B + C. A et B pouvant chacun être un complexe de protéines, le programme cherche la réaction de formation de AB pour déterminer les produits de la réaction de clivage. Chez Lambda, RecA clive le dimère CI_2.

La dégradation d'ARN est identique à la dégradation de protéine : a + B → B. Ainsi par exemple la RNaseIII dégrade les transcrits inititiés en P_L ayant passé la première terminaison de transcription ou les transcrits portant *cII* sur lesquels le petit transcrit produit en P_{OOP} s'hybride.

3.2.2 Coopérativité

On considère maintenant que la régulation de transcription par une protéine P_i nécessite n_i copies de cette protéine. Le taux de transcription est postulé être une fonction de Hill généralisée :

$$\frac{k_l + \sum_i k_{o,i} K_i [P_i]^{n_i}}{1 + \sum_i K_i [P_i]^{n_i}}$$

Ce type de paramétrisation semble bien reproduire la dynamique observée expérimentalement [54, 55]. Lors de la création d'une régulation, un degré n_i est attribué à la protéine

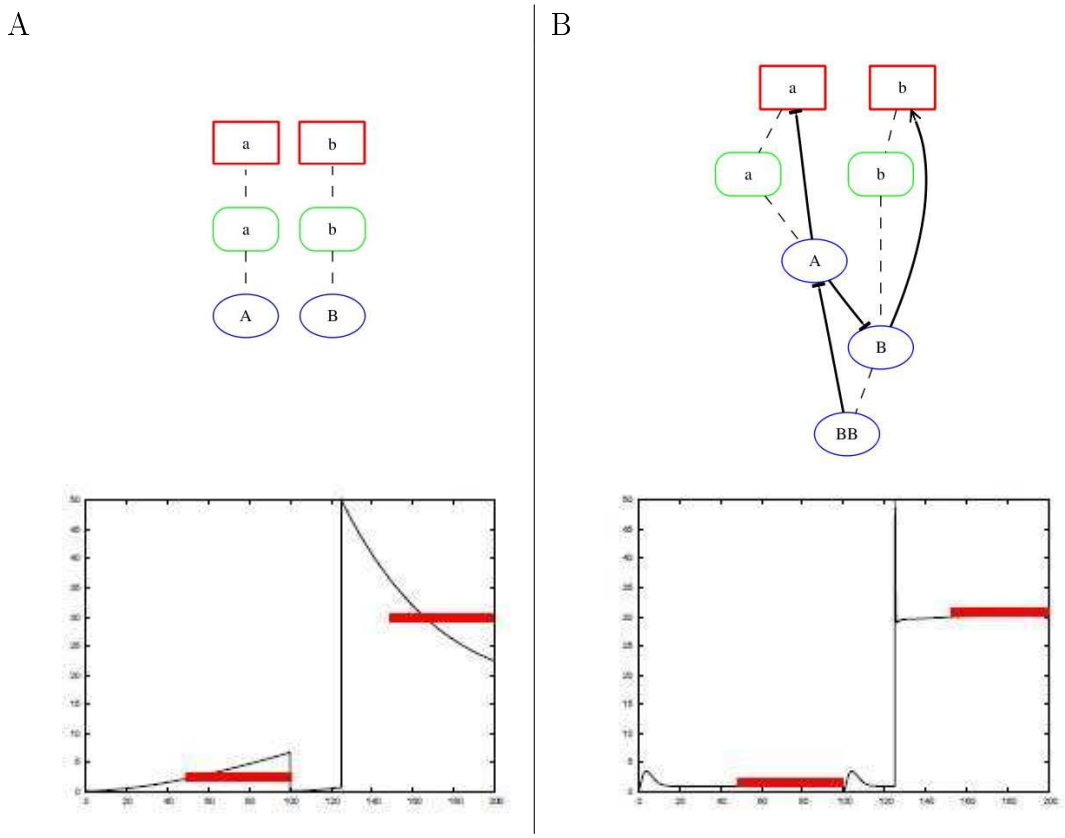


Fig. 3.1 – Exemple d'utilisation de *Genherite* dans sa version initiale : génération d'un réseau bistable. (A) Réseau de départ : deux gènes a et b sont transcrits en ARN a et b, eux-mêmes traduits en protéines A et B ; (B) réseau généré : A réprime la transcription du gène a et dégrade B, B active la transcription de b et forme un dimère qui dégrade A. En bas, le nombre de protéines A au cours du temps ($t = 0$ à 200). À $t = 100$, les conditions initiales sont de nouveau imposées, à $t = 125$ une valeur de $[A]$ est imposée. Le score est l'opposé du carré de la distance entre la courbe du nombre de A et la valeur cible indiquée par les barres rouges. En (A), le réseau converge vers un état stationnaire unique, en (B) les deux états stationnaires souhaités sont atteints.

régulatrice P_i. Par défaut il vaut 1, mais il peut-être muté lors de la procédure d'évolution. La coopérativité des dimères de CI s'attachant à l'ADN joue un rôle important dans la stabilité de la lysogénie [56, 35].

3.2.3 Opérons

Les opérons sont des objets constitués de gènes, promoteurs et terminaisons de transcription. Les promoteurs et les gènes sont orientés, mais pas les terminaisons.

Il y a, pour chaque promoteur, autant d'ARN que de terminaisons sous ce promoteur, plus un. Chaque ARN porte un ou plusieurs transcrits de gène. Un ARN est orienté, de même orientation que le promoteur sous lequel il est créé.

Les promoteurs sont régulés comme décrit précédemment. Les terminaisons sont régulées de la même manière que les promoteurs, si ce n'est que les taux de transcriptions sont remplacés par des taux de passage t, compris entre 0 et 1 (la probabilité que la polymérase s'arrête à la terminaison est alors $1-t$).

Le taux de transcription devient ainsi (j et k repèrent la position dans l'opéron du promoteur et de la terminaison qui définissent l'ARN : la transcription débute au promoteur j et s'arrête à la terminaison k) :

$$k^{(j)}\left(1-t^{(k)}\right)\prod_{j<i<k} t^{(i)}\ ,\ \text{où}\ \left\{\begin{array}{rcl} k^{(j)} & = & \frac{k_l^{(j)}+\sum_{i_j} k_{o,i_j}^{(j)} K_{i_j}^{(j)}[P_{i_j}]^{n_{i_j}^{(j)}}}{1+\sum_{i_j} K_{i_j}^{(j)}[P_{i_j}]^{n_{i_j}^{(j)}}} \\ t^{(p)} & = & \frac{t_l^{(p)}+\sum_{i_p} t_{o,i_p}^{(p)} K_{i_p}^{(p)}[P_{i_p}]^{n_{i_p}^{(p)}}}{1+\sum_{i_p} K_{i_p}^{(p)}[P_{i_p}]^{n_{i_p}^{(p)}}} \end{array}\right.$$

Remarque : il n'y a pas de terminaison en fin d'opéron : pour l'ARN allant jusqu'au bout, k est la position du dernier élément de l'opéron dans cette direction et $t^{(k)}=0$.

À chaque ARN sont associées des réactions de traduction, une pour chaque gène de même orientation que lui dont l'ARN porte le transcrit. Le taux de traduction n'est fonction que du gène : quels que soient les ARN qui en porte un transcrit, le taux de traduction du produit d'un gène est le même.

De nouvelles mutations sont introduites :
- créer ou supprimer un promoteur, une terminaison ;
- déplacer un promoteur, un gène, une terminaison ;
- fusionner deux opérons, avec ou sans changement d'orientation ;
- scinder un opéron.

Des règles empêchent la création d'opérons inutilement compliqués ou auxquels seraient associés des ARN vides. Ainsi, un gène devra toujours être sous un promoteur (de même orientation), il devra toujours y avoir un gène entre deux terminaisons, il ne pourra pas y avoir de terminaison en extrémité d'opéron, une terminaison devra toujours être en aval d'un promoteur (pas de terminaison entre deux promoteurs tête-bêche), il devra toujours y avoir au moins un gène ou une terminaison entre deux promoteurs de même orientation.

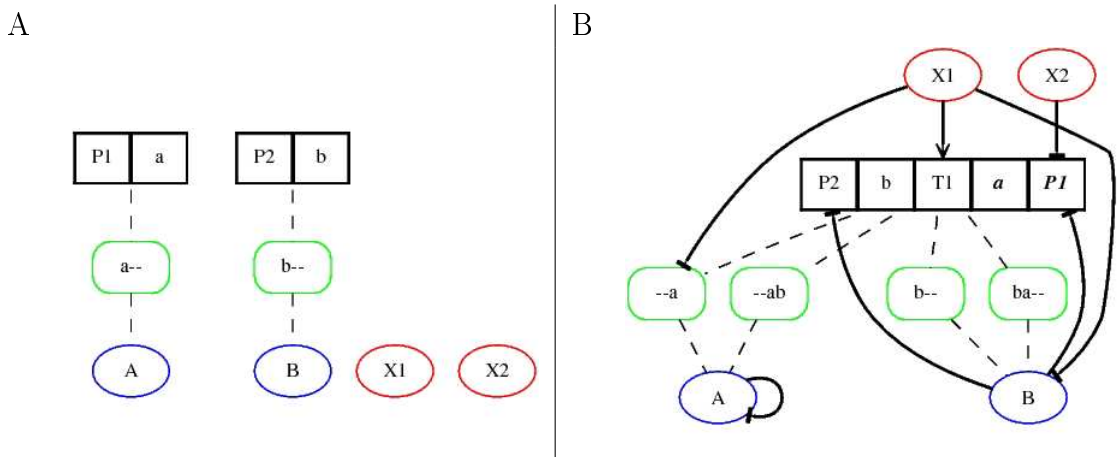


Fig. 3.2 – Exemple de génération de réseau avec organisation en opéron. (A) Réseau de départ, identique à celui de la figure 3.1, avec les promoteurs indiqués et deux protéines extérieures X1 et X2 ; (B) un réseau généré : les deux opérons de départ ont fusionné, en orientations inverses (P2-b : direct, *P1-a* : inverse) et une terminaison de transcription a été insérée ; quatre ARN sont produits, les tirets indiquant leur orientation. Seuls les transcrits de gènes ayant la même orientation que l'ARN sont traduits ; ainsi, seule B est produite à partir de l'ARN « ba– ». A catalyse sa propre dégradation, B inhibe les deux promoteurs, X2 inhibe P1 et X1 dégrade l'ARN a et facilite le passage à travers T1, activant ainsi la production des ARN ab et ba.

La figure 3.2 présente un exemple de réseau pouvant ainsi être généré. Des extraits de code présentant l'implémentation des opérons sont reproduits en annexe A.5.3.

3.2.4 Délais

Pour simplifier, je n'ai introduit, pour tous les réseaux, que deux délais : un « temps de transcription » d'un gène (le temps de transcription d'un ARN est alors égal à ce temps multiplié par le nombre de gènes dont il contient un transcrit) et un « temps de traduction » nécessaire à la polymérisation, la maturation et diffusion vers son site d'action ou son co-réactif d'une protéine. Toutes les autres réactions et les régulations se passent à temps égal.

Remarque : l'existence de délais rend compliquée l'utilisation d'algorithmes d'intégration à pas variable et/ou efficaces (Runge-Kutta par exemple) : le but étant de trouver des réseaux se distinguant par un comportement qualitativement particulier, intégrer « à la main » (méthode des rectangles) est suffisant.

Dans [57] les auteurs considèrent les délais comme des paramètres mutables, au même titre que les constantes cinétiques.

3.3 Perspectives

3.3.1 Comparaison avec Lambda

Le but principal du programme est de trouver des réseaux pouvant assumer une certaine fonction, et les comparer à des réseaux connus [53], les identifier dans des réseaux plus grands [58], suggérer des processus génétiques capables de produire un phénotype observé [57], etc.

Dans la perspective d'une comparaison des réseaux générés avec Lambda, trois processus mériteraient d'être implémentés : deux polymérases transcrivant une même portion d'ADN en sens inverse peuvent se gêner : le taux de transcription initiée à un promoteur est plus faible s'il existe un promoteur d'orientation inverse en aval de lui (cela pourrait notamment jouer un rôle dans la décision lyse/lysogénie chez Lambda [11, 37]) ; deuxièmement, l'ARN-polymérase peut faire une pause au passage de terminaisons de transcription, ce qui induit un délai supplémentaire ; enfin, la transcription d'un brin d'ADN peut diminuer l'occupation d'un site par des facteurs de transcription sur le brin complémentaire [37].

Les réseaux devront être soumis, au cours de la procédure d'évolution, à une pression de sélection qui corresponde à la fonction apparemment remplie par le réseau de régulation de Lambda. Une telle pression pourrait s'écrire :
- les réseaux doivent avoir au moins deux gènes a et b ;
- *bistabilité* : il doit exister deux états stationnaires, 1 et 2 : dans l'état 1, $[A]$ doit être élevé (100 par exemple) et $[B]$ proche de 0, et inversement dans l'état 2 ;
- *engagement* : s'il arrive que $[A]$ (resp. $[B]$) dépasse cette valeur élevée, alors l'état stationnaire doit être 1 (resp. 2) ;
- *biais* : soit une protéine X1 (équivalent de HflB/FtsH) extérieure au réseau, si son nombre à l'instant initial est inférieur (resp. supérieur) à un seuil donné, alors l'état stationnaire doit être 1 (resp. 2) ;
- *multiplicité d'infection* : quelles que soient les conditions, un réseau en plus de quatre copies doit atteindre un état stationnaire où $[B]$ est supérieur ou égal à sa valeur dans l'état 2 [4] ;
- *immunité à la surinfection* : soit le réseau dans l'état 2, l'ajout d'une copie de ce réseau doit conduire à un état stationnaire où $[B]$ est supérieur ou égal à sa valeur dans l'état 2 ;
- *induction de la lyse* : soit une protéine X2 (équivalent de RecA) extérieure au réseau, au-delà d'un seuil de $[X2]$, l'état 2 doit basculer vers l'état 1.

On peut ajouter une contrainte cinétique : que la valeur élevée de $[A]$ ou $[B]$ ne soit atteinte qu'après plusieurs minutes, pour que la décision dépendent peu de fluctuations rapides du nombre de facteur d'hôte X1.

4. En considérant qu'il n'y a pas de titre des facteurs de transcription par leurs sites de fixation, multiplier le nombre de copies du réseau revient à multiplier les taux de transcription effectifs.

3.3.2 Estimer les constantes cinétiques d'un réseau de topologie connue

Le programme peut aussi être utilisé de façon plus limitée : en partant d'un réseau dont on connaît la topologie, en ne considérant pas les mutations susceptibles de la changer et en utilisant les mesures expérimentales comme fonction de score, on peut espérer en déterminer les constantes cinétiques.

3.3.3 Comparer l'organisation du génome entre prokaryotes et eukaryotes : le rôle des opérons

Les opérons sont très peu fréquents chez les eukaryotes [59], alors qu'ils sont très répandus chez les prokaryotes. Deux utilisations du programmes peuvent permettre d'aborder cette question : 1. considérer des évolutions avec ou sans possibilité que les gènes s'organisent en opérons, en gardant la même pression de sélection ; 2. laisser la possibilité de former des opérons, mais ajouter des contraintes spécifiques à chaque domaine (existence d'un noyau pour les eukaryotes, transfert horizontal plus fréquent chez les prokaryotes, par exemple).

Conclusion

Le réseau de régulation génétique responsable de la décision lyse/lysogénie du bactériophage Lambda est un bon modèle de dynamique d'expression génétique. Nous avons développé et validé des méthodes qui permettront de mieux le décrire, et plus généralement d'obtenir un degré de connaissance inédit d'un tel réseau.

En insérant des gènes codant pour des protéines fluorescentes dans le génome de Lambda, nous avons pu suivre l'expression de deux gènes clés du réseau. En particulier, il a ainsi été possible de distinguer, au sein d'une population, les bactéries ayant suivi la voie lytique de celles ayant suivi la voie lysogénique. L'utilisation d'un répresseur thermosensible a permis de simuler l'infection et de mesurer les corrélations d'expression de ces deux gènes au cours de la décision. Nous avons pu suivre les expressions transitoires des gènes *cro* et *cI* vers leurs états stationnaires. Le système développé permettra d'étudier en détail l'agencement temporel des cascades de régulation conduisant à la lyse ou à la lysogénie, d'apprécier le rôle des fluctuations stochastiques, de mieux comprendre la stabilité de la lysogénie et l'induction de la lyse.

Les résultats obtenus pourront être améliorés en augmentant le nombre de mesures réalisées, en contrôlant mieux le protocole de dénaturation des répresseurs et les conditions de croissance des bactéries au cours des mesures. Ensuite, nous pourrons répéter les expériences sur d'autres constructions (d'autres couples de gènes étant marqués) pour obtenir une description quantitative riche de ce réseau. Un système microfluidique de contrôle du milieu de croissance au cours de l'observation constituerait une amélioration importante.

Les apports au programme *Genherite* permettront non seulement de comparer des réseaux générés sur ordinateur avec Lambda, mais aussi peut-être de mieux analyser des données de cinétique de réseaux de régulation génétique et de comparer l'organisation des génomes prokaryotes et eukaryotes.

La démarche suivie au cours de ce travail pourra être généralisée à d'autres systèmes où l'on pense que la dynamique d'expression, et en particulier les fluctuations stochastiques, peut jouer un rôle important.

Annexe A

Matériels et méthodes

A.1 Biologie moléculaire

A.1.1 Constructions

Les trois constructions décrites au paragraphe 2.2 ont été réalisées par Geneart [1], sur indication des séquences, et livrées clonées dans des vecteurs d'origine de réplication ColE1 (pMK et pCR4). Des séquences de quarante bases d'homologie avec le chromosome de *E. coli* (autour du site *attB* d'insertion naturel de Lambda) [2] et des sites de restriction NotI et SacI ont été ajoutés de part et d'autre des constructions. Ces plasmides ont été séquencés avant d'être livrés.

A.1.2 Clonage

Extraction, restriction, ligation et transformation

Le kit MaxiPrep Qiagen a été utilisé pour extraire le plasmide pZC320 d'une culture de bactéries.

Le plasmide pZC320 et les plasmides portant les constructions ont été digérés par NotI et SacI. Les produits de digestion ont été purifiés sur gel, puis le pZC320 ouvert et les constructions ont été mélangés en proportions identiques (mélange équimolaire) et ligués par l'ADN-ligase du phage T4.

Le produit de ligation a été directement électroporé dans des bactéries TOP10 (voir le paragraphe A.2.1). Après électroporation et incubation 2h dans du LB à 37 °C, les bactéries ont été étalées sur du LB-agar supplémenté en streptomycine et ampicilline [3], et couvert d'IPTG et X-gal. Les sites de restriction du pZC320 se trouvant dans le gène *lacZα* (supprimé du génome de TOP10), les colonies portant le plasmide intact apparaissent

1. http ://www.geneart.com
2. Ces séquences ont été ajoutées dans la perspective d'insérer ces constructions dans le génome de la bactérie par recombinaison ; aucun recombinant n'a cependant pu être obtenu.
3. La souche TOP10 est résistante à la streptomycine et le plasmide pZC320 porte la résistance à l'ampicilline : ainsi seules les bactéries transformées pousseront après l'électroporation.

bleues, celles portant un plasmide contenant un fragment cloné appraissant blanches (test de β-galactosidase).

Vérifications : PCR, séquençage, fluorescence

Pour vérifier que chaque construction a correctement été clonée dans le pZC320, quatre PCR ont été réalisées sur des bactéries issues des colonies apparaissant blanches au test de β-galactosidase avec des amorces homologues à des régions de Lambda. Les régions amplifiées, autour de la terminaison *TeT7* (de la région d'homologue à *E. coli* à *cIII*), de *cIII* à *pRM*, de *cI* à *pRE*, de *cro* à la région d'homologue à *E. coli* à droite de *cII*, avaient la longueur attentue.

Pour les construction pZC-Lf2 et pZC-Lf3, des séquençages on confirmé l'intégrité des fragments clonés.

Enfin, les trois souches apparaissent rouges ou vertes en fluorescence, suivant les conditions de culture.

A.2 Microbiologie

A.2.1 Souche TOP10

La souche TOP10 de *E. coli*, commercialisée par Invitrogen, a été utilisée tout au long de cette étude. Son génotype est [4] :

F- mcrA Δ(mrr-hsdRMS-mcrBC) ϕ80lacZΔM15 ΔlacX74 nupG recA1 araD139 Δ(ara-leu)7697 galE15 galK16 rpsL(StrR) endA1 λ-

La mutation *recA1* rend RecA inactif, ce qui offre une meilleure stabilité des plasmides. Il faut en tenir compte si l'on veut étudier la stabilité de la lysogénie, beaucoup plus grande en l'absence de RecA.

A.2.2 Antibiotiques et milieux

Antibiotiques

Toutes les bactéries utilisées, dérivées de TOP10, sont résistantes à la streptomycine. Le plasmide pZC320 porte la résistance à l'ampicilline et les plasmides pMK et pCR4 portent la résistance à la kanamycine. Suivant [60], la streptomycine a été utilisée à une concentration de 10 μg/ml, l'ampicilline 20 μg/ml et la kanamycine 15 μg/ml (dans ce dernier cas, la concentration recommandée varie de 10 à 50 μg/ml suivant le nombre de copies du plasmide sélectionné). Les solutions d'antibiotiques sont stérilisées par filtrage (0,22 μm).

4. voir par exemple http ://openwetware.org/wiki/E._coli_genotypes

Milieu « Luria-Bertani » (LB)

Ce milieu riche favorise la lyse lors de l'infection d'une bactérie par Lambda. Il a été obtenu par dissolution de la préparation « LB-Broth, high salt » de Fluka dans de l'eau distillée (25g/l, de façon à obtenir des concentrations finales d'hydrolysat enzymatique de caséine de 10g/l, d'extrait de levure de 5g/l et de chlorure de sodium de 10 g/l).

Il est stérilisé par passage à l'autoclave avant utilisation.

Milieu minimal M9 (MM)

Ce milieu pauvre favorise la lysogénie lors de l'infection d'une bactérie par Lambda. Sa préparation suit le protocole de [60] : pour un litre, 750 ml d'eau distillée, 200 ml de sels M9 à 5X, 2 ml de $MgSO_4$ à 1 M, 20ml d'une solution de glucose à 20%, 0,1 ml de $CaCl_2$ à 1 M.

Il a été supplémenté en L-Leucine (utilisée à une concentration finale de 100 μg/ml), la souche TOP10 étant incapable de la synthétiser.

La solution de sels M9 à 5X est préparée comme suit : dans un litre d'eau distillée sont dissous 64 g de $Na_2HPO_4{\cdot}7H_2O$, 15 g de KH_2PO_4, 2,5 g de NaCl, 5,0 g de NH_4Cl.

Les solutions de M9, $MgSO_4$ et $CaCl_2$ sont stérilisées séparément par passage à l'autoclave. Les solutions de glucose et leucine sont stérilisées par filtrage (0,22 μm).

Phosphate-buffered Saline (PBS)

La solution de PBS est un tampon qui évite aux bactéries d'être détruites par pression osmotique et maintient un pH adapté. Il ne contient pas de nutriment, si bien que les bactéries ne peuvent y croître et n'y produisent pas (ou peu) de protéines.

A.2.3 Préparation des échantillons

170 μl d'agar (10 g/l) supplémenté en LB ou en MM fondu (60 °C) est versé sur une lame de microscope nettoyée à l'éthanol. Un cadre d'hybridation [5] a été préalablement collé sur la lame pour assurer que l'agar ne coule pas et qu'il garde une épaisseur constante. Une deuxième lame est immédiatement posée sur la cadre, de telle sorte que l'agar s'étale et qu'il ait une surface plane. 30 min plus tard, la deuxième lame est retirée et 4 μl de culture de bactéries sont étalées sur l'agar et laissés à sécher 3 min sous aération de hotte. Enfin, une lamelle est collée sur le tout, tenue par le cadre ; l'agar portant la culture de bactéries est suffisamment humide pour mouiller la lamelle.

5. Il s'agit d'une bande dessinant un rectangle et adhésive sur chaque face ; référence : Cadre pour hybridation in-situ Gene Frame 125 μl ABgene.

A.3 Imagerie

A.3.1 Montage

Une microscope droit Olympus BX51WI, sur lequel sont montés une caméra MicroMax (Princeton Instruments), un objectif 100X à immersion et une lentille de grandissement x2 est utilisé pour imager les bactéries. Des images de 512×512 pixels sont produites.

Une platine motorisée, pilotée par *joystick* ou par ordinateur, permet de régler la position de l'échantillon sous le microscope.

De l'air chauffé circule dans une cage en plexiglas entourant le microscope et la platine, de manière à maintenir l'échantillon à la température désirée au cours de l'observation.

A.3.2 Éclairage

Un système de contraste de phase est utilisé lors de l'éclairage en lumière blanche, notamment lors de la mise au point (voir le paragraphe suivant).

Une lampe à mercure (100 W) est utilisée pour l'éclairage de fluorescence. Bien que le mercure n'ait pas de raie d'émission à la longueur d'onde d'excitation de mCherry, l'intensité « basale » est suffisante pour qu'on détecte très bien les niveaux de fluorescence rouge de pZC-Lf1 en lysogénie.

Un module de changement de filtres motorisé est installé sur le microscope. Des cubes de filtres adaptés à EGFP [6] et mCherry [7] sont montés. Des filtres de densité neutre ont été ajoutés à chacun des cubes de manière à réduire l'intensité d'excitation : un filtre « ND03 » sur le cube EGFP, un filtre « ND10 » sur le cube mCherry [8]. À chaque image, mCherry est excitée pendant 1 seconde, EGFP pendant 3. Une image est prise toutes les 5 minutes.

Une lame de verre a été ajoutée sur le trajet de la lumière de la lampe à mercure pour s'assurer que les ultra-violets sont bien coupés.

A.3.3 Acquisition des images

Le pilotage du microscope, de la platine et de la tourelle de changement de filtre se fait par un programme Labview que nous avons écrit (l'essentiel a été fait par Jérôme Wong Ng).

À chaque nouvelle image (toutes les 5 minutes) la mise au point est faite, l'échantillon pouvant bouger légèrement au cours de l'observation. L'objectif est déplacé de 30 pas de 4 μm autour de sa position à l'image précédente ; à chaque pas, le contraste de l'image obtenue est calculé, l'image nette ayant le contraste le plus fort. Cette image est écrite sur le disque, ainsi que les images juste avant et juste après le point : une image légèrement

6. Jeu de filtres GFP-3035B-000 (GFP BrightLine Filter Set) de Semrock.
7. Jeu de filtres TXRED-4040B-000 (Texas Red BrightLine Filter Set) de Semrock.
8. Un filtre ND03 transmet $10^{-0,3} = 50\%$ de l'intensité lumineuse reçue, un filtre ND10 $10^{-1,0} = 10\%$.

floue est utilisée par *Schnitzcells* (le programme d'analyse des images, voir le paragraphe suivant) pour détourer les bactéries.

Après la mise au point, une image de fluorescence dans chaque canal, rouge et vert, est prise (éclairage par la lampe à mercure, cube adapté positionné sur le trajet optique).

Deux champs éloignés sont filmés sur chaque échantillon.

A.4 Analyse

A.4.1 Extraction de la fluorescence de bactéries individuelles

Le programme Matlab *Scnitzcells*, développé dans le laboratoire de Michael Elowitz, a été utilisé pour extraire les données de croissance et de fluorescence des bactéries imagées [9].

Par seuillage des images de contraste de phase, il génère des masques identifiant des bactéries individuelles. Puis le film est analysé de manière à reconstituer le lignage des cellules. Chaque étape peut être corrigée manuellement.

Diverses informations sur les bactéries, en particulier leur longueur, largeur, fluorescence dans chaque canal à chaque image, leurs mère et filles, sont extraites et compilées dans une structure. L'intensité de fluorescence de chaque bactérie est corrigée en lui soustrayant à chaque image la valeur médiane du fond (bord de l'image). Le volume est estimé en supposant que les bactéries sont cylindriques, de diamètre leur largeur.

A.4.2 Correction des données

Pour ne prendre en compte que les protéines EGFP produites au cours de l'observation, la valeur initiale d'intensité de fluorescence verte d'une bactérie est retranchée à toutes ses descendantes, multipliée par 1/2 à chaque division. Pour les protéines mCherry, il faut aussi prendre en compte l'extinction de fluorescence. On peut, en utilisant la forme d'extinction de fluorescence décrite au paragraphe 2.3.1, calculer à chaque pas de temps t l'intensité de fluorescence dI issue des protéines produites entre ce pas et le précédent : $dI(t)$ est l'intensité $I(t)$ mesurée à laquelle est retranchée l'intensité des protéines nouvelles produites à chaque pas de temps précédent dans la lignée de la bactérie considérée, divisée par 2 à chaque division et corrigée pour l'extinction :

$$dI(t) \quad = \quad I(t) - \sum_{t_1<t} k_{t,t_1}\, I_{\text{bleach}}(t-t_1)\, dI(t_1)$$

où k_{t,t_1} vaut 1/2 à la puissance le nombre de divisions dans la lignée entre t_1 et t, I_{bleach} est la courbe d'extinction de fluorescence attendue dans ces conditions (éclairage de 1 s toutes les 5 min, filtre de densité neutre ND10) normée à 1 en 0 et on a posé $dI(0) = 0$.

Enfin, pour retrouver une intensité corrigée $I_{\text{corr}}(t)$ proportionnelle au nombre de protéines mCherry dans la bactérie considérée produites depuis le début de l'observation, il

9. Il peut être téléchargé à l'adresse http ://www.elowitz.caltech.edu/software

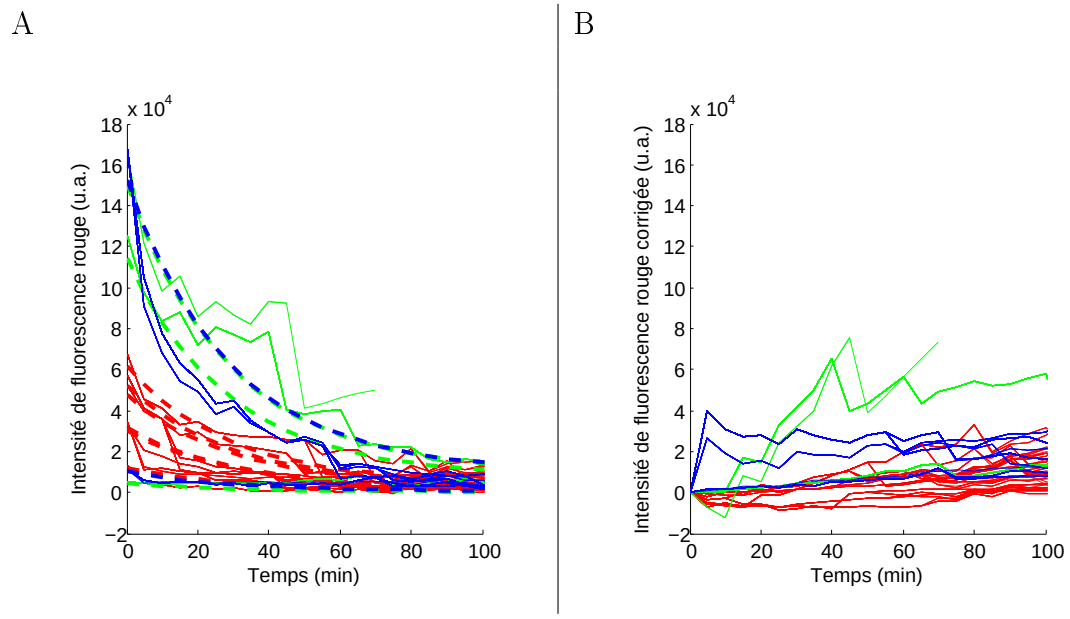


Fig. A.1 – Intensités de fluorescence rouge de bactéries pZC-Lf1 sur du LB à 37 ° C ; chaque courbe pleine correspond à une bactérie, chaque couleur à un film différent. (A) Intensité brutes, et, en pointillés, courbes d'extinction de fluorescence attendues pour chaque bactérie à l'instant initial (sans prendre en compte la dilution lors des divisions). (B) Intensité de fluorescence corrigée pour l'extinction de fluorescence et la dilution lors des divisions. Les intensités n'ont pas été divisées par le volume.

faut prendre la somme des intensités de toutes les nouvelles protéines produites à chaque pas de temps précédent, divisées par 2 à chaque division :

$$I_{\mathrm{corr}}(t) \quad = \quad \sum_{t_1<t} k_{t,t_1}\, dI(t_1)$$

La figure A.1 présente les données de fluorescence rouge avant et après correction, pour des bactéries pZC-Lf1 sur du LB à 37 ° C.

Après ces corrections, chaque intensité est divisée, à chaque pas de temps, par le volume de la bactérie.

Un extrait du script Matlab utilisé est reproduit ci-dessous. Les grandeurs calculées sont `IntV(j,k,i)` et `IntR(j,k,i)`, les intensités verte et rouge corrigées et divisées par le volume de la cellule `i`, au pas de temps `k`, dans l'expérience `j`.

```
iS = 0;
for j=1:Nexp
    nMins=zeros();
%     NschnitzExp(j) est le nombre total de bacteries dans l'experience j
    if(j > 1) iS = iS + NschnitzExp(j-1); end
```

```
%    data(j) est le chemin des donnees (.mat) de l'experience j
    load(char(data(j)));

%     IntVinit est la part de la fluorescence verte de chaque cellule heritee de la premiere
    IntVinit = zeros(1,length(schnitzcells));

    for k=1:length(p.trackRange)
        for i=1:length(schnitzcells)
            if (schnitzcells(1,i).approved==1 ...
                    && (schnitzcells(1,i).D > 0 || schnitzcells(1,i).P > 0))
                for l=1:length(schnitzcells(1,i).frames)
                   if (schnitzcells(1,i).frames(l)==k)
                      if (k==1)
                          m = m+1;
%                           famille(i) est la famille a laquelle appartient la cellule i
                          famille(i+iS) = m;
                          NImFamille(m) = length(p.trackRange);
                          IntVinit(i) = schnitzcells(1,i).FG(1);
                      elseif(l==1)
                          IntVinit(i) = IntVinit(schnitzcells(1,i).P)/2;
                          famille(i+iS) = famille(schnitzcells(1,i).P+iS);
                      end

                      IntV(j,k,i) = (schnitzcells(1,i).FG(l) - IntVinit(i)) ...
                          /schnitzcells(1,i).volume(l);

%                       dIR(j,k,i) est la fluorescence provenant des proteines produites
%                       entre k et k-1 dans la cellule i de l'experience j
                      dIR(j,k,i) = schnitzcells(1,i).FR(l);
                      div = 1;
                      i1 = i;
                      if(k>1)
                          for k1=(k-1):-1:1
                              if (schnitzcells(1,i).P>0 ...
                                 && k1 == schnitzcells(1,schnitzcells(1,i).P).frames(end))
                                  i1 = schnitzcells(1,i1).P;
                                  div = div/2;
                              end
                              dIR(j,k,i) = dIR(j,k,i) ...
                                  - div*Ifit(k-k1)*IntRNew(j,k1,i1);
                          end
                      end

                      div = 1;
                      i1 = i;
```

```
                    if (k==1)
                        IntR(j,k,i) = 0;
                    else
                        for k1=k:-1:2
                            IntR(j,k,i) = IntR(j,k,i) ...
                                + div*IntRNew(j,k1,i1);
                            if (k1==schnitzcells(1,i1).frames(1))
                                i1 = schnitzcells(1,i1).P;
                                div = div/2;
                            end
                        end
                    end
                    IntR(j,k,i) = IntR(j,k,i)/schnitzcells(1,i).volume(l);
                end
            end
        end
    end
end
```

A.5 *Genherite* : extraits de classes et constructeurs

Genherite, écrit en C++, tire parti de la programmation orientée objet. Je présente dans la suite les principales classes que j'ai introduites ou redéfinies.

A.5.1 Les blocs : gènes, promoteurs, terminaisons

Les Blocs sont les constituants des opérons : gènes, promoteurs ou terminaisons. Ils sont orientés, et par commodité les constantes de traduction (pour les gènes), de transcription libre (promoteurs) et de passage (terminaisons) y sont définies.

```
class Bloc
{//élements d'un opéron : promoteur, gène ou terminaison de transcription
   public:
      int nombre;
      double constante;// constante de traduction (gène),
                       //transcription libre (promoteur),
                       //de "terminaison" libre (terminaison), en fait de passage
                       //(dans ce dernier cas : doit être < 1
                       //taux de terminaison libre :
                       //0 = pas de passage de la polymérase / 1 = pas de terminaison)
      Bloc *copie;
      string label;
      string type; // "p"=promoteur / "g"=gène / "t"=terminaison
      int orient; // 1 = direct / -1 = reverse ...
                  //les terminaisons sont symétriques : orient = 0
```

```
};

class Gene : public Bloc
{
   public:
      Protein *prot;

      Gene(string lab,double c=frand2());
      void modifconstante(double ampli=defampli);
};
```

Trois classes en dépendent : Gene, Promoteur et Terminaison. À un gène est associée une protéine (son produit), créée avec lui, et une constante de traduction. À un promoteur est associé un ensemble de protéines régulatrices et pour chacune une constante d'équilibre, un taux de transcription « occupé » et un degré de Hill. La classe Terminaison est identique à Promoteur à la différence que les terminaisons ne sont pas orientées (orient=0) et que leur constante (taux de passage) est inférieure à 1.

```
Gene::Gene(string lab,double c)
{
   label=lab;
   constante=c;
   type="g";
   orient=1;
   prot=new Protein();
   prot->qtite=seuil_qtite_inf;
   prot->label=lab;
   char labp=lab[0];
   prot->label[0]=toupper(labp);
}

class Promoteur : public Bloc
{
   public:
      vector<Protein *> prot_reg; //protéines régulatrices
      vector<double> cte_ad; //rapports des constantes d'association et
                             //dissociation des prot régulatrices
      vector<double> cte_occ; //doit être < 1 / taux de terminaison du
                              //promoteur occupé par une prot régulatrice : idem
      vector<int> hill_degr; //exposant des concentrations dans la fonction
                             //de Hill ~ coopérativité

      Promoteur(double c=frand2());
      void modifconstante(double ampli=defampli);
      void modifcte_occ(double ampli=defampli);
```

```
      void modifcte_ad(double ampli=defampli);
      void modifhill_degr(double ampli=defampli);

};

Promoteur::Promoteur(double c)
{
  constante=c;
  type="p";
  orient=1;
}
```

A.5.2 ARN

Un Arn contient les transcrits orientés de Genes ; les Reactions (formées d'un nom, de deux réactifs et trois produits, éventuellement vides (NULL), et d'une constante de réaction) de traduction associées sont créées avec lui : pour chaque gène de même orientation que lui dont l'ARN porte un transcrit est associée une réaction produisant la protéine associée à ce gène.

La classe Arn est une sous-classe de la classe Espece, constituée d'un nom, d'une constante de dégradation et d'une quantité initiale.

```
class Arn : public Espece
{
   public:
      double cste;
      int orient; //orientation de l'ARN, la même que celle du promoteur sous
                  //lequel il est produit ; seuls les gènes de même
                  //orientation sont traduits.
      vector<Gene  *> genes;
      vector<Reaction *> traductions;

      Arn(int orient,vector<Gene  *> genes,double c=frand2(),double q=frand2());
      ~Arn();
      void modifdegrad(double ampli=defampli);
      void modifqtite(double ampli=defampli);
};

Arn::Arn(int ori,vector<Gene *> gens,double c,double q)
:genes(gens)
{
   orient=ori;
   cste=c;
   qtite=q;
   for(unsigned int i=0;i!=gens.size();i++){
```

```
    label+=gens[i]->label;
  }
  for(unsigned int i=0;i!=gens.size();i++){
    if(gens[i]->orient==this->orient){
      Reaction *react=new Reaction("traduction",this,NULL,
           this,gens[i]->prot,NULL,gens[i]->constante);
      traductions.push_back(react);
    }
  }
}
```

A.5.3 Opérons

Les Operons sont constitués de Blocs (gènes, promoteurs, terminaisons) ; les Arns associés sont créés avec lui. Les constantes de transcriptions effectives sont calculées au cours de l'intégration, ailleurs dans le programme.

```
class Operon : public Espece
{
  public:
     vector<Bloc *> blocs;
     vector<Arn *> arns;

     Operon(vector<Bloc *> blocs);
     ~Operon();
};

Operon::Operon(vector<Bloc *> bloks)
:blocs(bloks)
{
  qtite=1.0;

  vector<Bloc *>::iterator ibloc_p,ibloc_t,ibloc_g;
  vector<Bloc *>::reverse_iterator ribloc_t,ribloc_g;
  vector<Gene *> genes;
  Arn * arn;
  Bloc *termf=bloks[0];
  unsigned int numpromD=0;
  unsigned int numpromR=0;

  for(ibloc_p=bloks.begin();ibloc_p!=bloks.end();ibloc_p++){
    if((*ibloc_p)->type=="p"){
      for(unsigned int num=0;num<bloks.size();num++){
        if(bloks[num]==*ibloc_p) numpromD=num;
      }
      numpromR=bloks.size()-1-numpromD;
```

```
//Boucle sur les terminaisons qui sont en aval du promoteur
//(sauf juste après : pas d'ARN vide)
      switch((*ibloc_p)->orient){
        case 1://Pomoteur "direct"
          for(ibloc_t=ibloc_p+2;ibloc_t!=bloks.end();ibloc_t++){
            if((*ibloc_t)->type=="t"){
              for(ibloc_g=ibloc_p;ibloc_g!=ibloc_t;ibloc_g++){
                if((*ibloc_g)->type=="g"){
                  genes.push_back((Gene *) *ibloc_g);
                }
              }
              arn=new Arn((*ibloc_p)->orient,genes);
              arn->qtite=seuil_qtite_inf;
              arns.push_back(arn);
              genes.clear();
              termf=*ibloc_t;
            }
          }
//Prendre en compte les éventuels gènes après la dernière terminaison
          if(bloks[bloks.size()-2]==termf &&
             bloks[bloks.size()-1]->type!="g") break;
          for(ibloc_g=ibloc_p;ibloc_g!=bloks.end();ibloc_g++){
            if((*ibloc_g)->type=="g"){
                genes.push_back((Gene *) *ibloc_g);
            }
          }
          arn=new Arn((*ibloc_p)->orient,genes);
          arn->qtite=seuil_qtite_inf;
          arns.push_back(arn);
          genes.clear();
          break;
        case -1://Pomoteur "reverse"
          for(ribloc_t=bloks.rbegin()+numpromR+2;ribloc_t!=bloks.rend();
            ++ribloc_t){
            if((*ribloc_t)->type=="t"){
              for(ribloc_g=bloks.rbegin()+numpromR;ribloc_g!=ribloc_t;
                ++ribloc_g){
                if((*ribloc_g)->type=="g"){
                  genes.push_back((Gene *) *ribloc_g);
                }
              }
              arn=new Arn((*ibloc_p)->orient,genes);
              arn->qtite=seuil_qtite_inf;
              arns.push_back(arn);
              genes.clear();
```

```
              termf=*ribloc_t;
            }
          }
//Prendre en compte les éventuels gènes après la dernière terminaison
          if(bloks[1]==termf && bloks[0]->type!="g") break;
          for(ribloc_g=bloks.rbegin()+numpromR;ribloc_g!=bloks.rend();
            ++ribloc_g){
            if((*ribloc_g)->type=="g"){
                genes.push_back((Gene *) *ribloc_g);
            }
          }
          arn=new Arn((*ibloc_p)->orient,genes);
          arn->qtite=seuil_qtite_inf;
          arns.push_back(arn);
          genes.clear();
          break;
      }
    }
  }
}
```

Annexe B

Nombre de copies de plasmides

L'étude exposée dans cette section a été motivée par le travail de thèse de Jérôme Wong Ng [61], sous la direction de Didier Chatenay et Jérôme Robert, sur le nombre de copies de plasmides au sein d'une population monoclonale de bactéries.

N.B. Ce travail a été soumis à publication, un preprint *est reproduit en fin de manuscrit. Il contient notamment des résultats de simulations qui ne sont pas exposés ici.*

B.1 Introduction

Des bactéries portant le gène *egfp* sous le contrôle du promoteur inductible *pTacI* dans le chromosome ont été transformées avec des plasmides portant le gène *mOrange* sous le même promoteur *pTacI*[1]. Après une heure d'induction de la production des protéines fluorescentes, la production de protéines et la division sont bloquées. La fluorescence mesurée dans le vert (resp. orange) est proportionnelle au nombre de protéines EGFP (resp. mOrange) produites pendant l'induction.

On s'attend ainsi à ce que les niveaux de fluorescence mesurés reflètent les nombres moyens et les fluctuations de nombres de copies de plasmides et de chromosome. Je chercherai dans la suite à expliciter ce lien.

La production de protéines étant un phénomène bruité, il n'est *a priori* pas certain qu'il soit possible avec un tel dispositif de distinguer les fluctuations d'expression des fluctuations de nombre de copies. Un modèle simple (paragraphe suivant) permet de comprendre comment, en utilisant la référence sur le chromosome, on peut s'affranchir de ce problème (ou plus exactement le ramener à un problème de fluctuations de nombre de copies de chromosome). On fera des hypothèses reflétant le fait que la même construction est présente sur le chromosome et sur les plasmides : les taux moyens d'expression de *egfp* et *mOrange* sont les mêmes, les corrélations d'expression de deux copies de plasmide ou de chromo-

1. J'appellerai « nombre de copies de chromosome » (resp. de plasmide) le nombre de copies du gène *egfp* (resp. *mOrange*).

some différentes sont identiques aux corrélations entre une copie de plasmide et une copie de chromosome. On admettra qu'en moyenne l'expression de *egfp* et *mOrange* ne dépend pas du temps, ni, comme cela est vérifié, du nombre de plamsides.

On devra cependant prendre en compte d'autres contributions aux fluctuations de fluorescence : au cours de l'induction le chromosome et les plasmides sont répliqués, les bactéries se divisent et elles n'ont pas une distribution d'âges uniforme. Un modèle plus réaliste (paragraphe B.3) sera alors proposé. La démarche sera la même que dans le cas du modèle simple : exprimer les moments des nombres de protéines fluorescentes et essayer d'en extraire, en utilisant les mêmes hypothèses sur l'expression des gènes, ceux de nombres de copies de plasmide et de chromosome. J'introduirai deux hypothèses alternatives sur les temps de corrélation d'expression. Il ne sera plus possible d'obtenir des formules explicites, mais on pourra estimer les grandeurs cherchées en bornant des termes liés à la façon dont les plasmides et le chromosome se répliquent et se répartissent à la division.

Les calculs sont présentés aux pragraphes B.4 à B.7. Les résultats sont résumés et discutés en B.8, puis les conclusions tirées en B.9.

B.2 Modèle simple

Supposons qu'au cours de l'induction, les bactéries ne croissent pas, que les plasmides et le chromosome ne se répliquent pas, que la production de protéines ne dépend pas du temps[2] et que la distribution des âges est uniforme.

En notant P_a^i la contribution de la copie i du gène a ($a = O$ ou V pour les gènes *mOrange* ou *egfp*) au nombre total de protéines P_a à la fin de l'induction dans une bactérie et n_a le nombre de copies du gène a, on peut écrire :

$$P_a = \sum_{i=1}^{n_a} P_a^i$$

La moyenne de P_a s'écrit alors :

$$\langle P_a \rangle = \sum_{n_a} \sum_{i=1}^{n_a} \sum_{P_a^i} p(n_a, P_a^i)\, P_a^i$$

On peut supposer que la distribution du nombre de protéines produites par chaque copie ne dépend ni de la copie considérée, ni du nombre de copies. Soit :

$$\begin{aligned} \langle P_a \rangle &= \sum_{n_a} p(n_a)\, n_a \sum_{P_a^1} p(P_a^1)\, P_a^1 \\ &= \langle n_a \rangle \langle P_a^1 \rangle \end{aligned}$$

2. Cette hypothèse n'est pas nécessaire, supposer que la production ne dépend en moyenne pas du temps suffirait, mais elle permet d'alléger les notations.

On peut de plus supposer que le nombre moyen de protéines produites au cours de l'induction par une copie de gène ne dépend pas du type de gène ($\langle P_O^1\rangle = \langle P_V^1\rangle = \langle P^1\rangle$). On obtient alors :

$$\boxed{\frac{\langle n_O\rangle}{\langle n_V\rangle} = \frac{\langle P_O\rangle}{\langle P_V\rangle}}$$

Les moments d'ordre 2 s'écrivent :

$$\langle P_a P_b\rangle = \sum_{n_a,n_b}\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}\sum_{P_a^i,P_b^j} p(n_a,n_b,P_a^i,P_b^j)\,P_a^i P_b^j$$

Dans le cas de deux types de gènes différents, on peut supposer que la corrélation ne dépend pas des deux copies i et j de chacun des gènes, ni de leur nombre. Alors :

$$\begin{aligned}\langle P_O P_V\rangle &= \sum_{n_O,n_V} p(n_O,n_V)\,n_O n_V \sum_{P_O^1,P_V^1} p(P_O^1,P_V^1)\,P_O^1 P_V^1\\ &= \langle n_O n_V\rangle\langle P_O^1 P_V^1\rangle\end{aligned}$$

Dans le cas d'un même type de gène, on peut supposer que deux copies différentes corrèlent comme deux copies de gènes de types différents ($\langle P_a^i P_a^j\rangle = \langle P_O^1 P_V^1\rangle,\ \forall i \neq j$) et que l'autocorrélation d'une copie ne dépend pas de la copie considérée (numéro ou type : $\langle (P_a^i)^2\rangle = \langle (P^1)^2\rangle,\ \forall a,i$). Alors :

$$\langle P_a^2\rangle = \langle n_a\rangle\langle (P^1)^2\rangle + \langle n_a(n_a-1)\rangle\langle P_O^1 P_V^1\rangle$$

Après élimination des coefficients liés à l'expression génétique et remplacement de $\langle n_O\rangle/\langle n_V\rangle$ par $\langle P_O\rangle/\langle P_V\rangle$:

$$\boxed{\langle n_O^2\rangle = \frac{\langle P_O\rangle}{\langle P_V\rangle}\langle n_V^2\rangle + \frac{1}{\langle P_O P_V\rangle}\left(\langle P_O^2\rangle - \frac{\langle P_O\rangle}{\langle P_V\rangle}\langle P_V^2\rangle\right)\langle n_O n_V\rangle}$$

La réplication du chromosome est très bien contrôlée [62]. On peut ainsi supposer que la variance du nombre de copies chromosome est nulle ($\langle n_V^2\rangle \simeq \langle n_V\rangle^2$) et que les nombres de copies de plasmide et de chromosome sont décorrélés ($\langle n_O n_V\rangle \simeq \langle n_O\rangle\langle n_V\rangle$). On ne connaît pas le nombre moyen de chromosome, mais il n'apparaît pas dans l'expression du bruit de nombre de copies de plasmides η :

$$\boxed{\eta^2 := \frac{\langle n_O^2\rangle - \langle n_O\rangle^2}{\langle n_O\rangle^2} = \frac{\langle P_V\rangle}{\langle P_O\rangle} + \frac{1}{\langle P_O P_V\rangle}\left(\frac{\langle P_V\rangle}{\langle P_O\rangle}\langle P_O^2\rangle - \langle P_V^2\rangle\right) - 1}$$

Les valeurs obtenues expérimentalement pour η sont présentées dans le tableau B.2 du paragraphe B.8.2.

B.3 Modèle réaliste

Je présente dans la suite un modèle plus complet de la production des protéines fluorescentes, qui tient compte de la distribution des âges des bactéries, de leurs divisions et de la variation du nombre de copies de plasmide et de chromosome au cours de l'induction. Je supposerai qu'il n'y a pas de perte de plasmides et qu'ils ne constituent pas un « poids métabolique » (la croissance des bactéries n'est pas affectée par la présence de plasmides ; ces hypothèses sont vérifiées par l'expérience, [61]), que toutes les bactéries ont le même temps de division T [3] et que les protéines d'une cellule mère se répartissent aléatoirement et symétriquement entre les cellules filles, de manière indépendante à chaque génération.

Je ferai de plus l'hypothèse que le temps d'induction est un multiple du temps de division : toutes les bactéries ont le même temps de division $T = 20$ min et toutes les lignées sont induites pendant exactement une heure, si bien que pendant l'induction, il y a exactement trois divisions [4]. Cette hypothèse est discutée en conclusion B.9.

Soit une lignée particulière de bactéries induites, c'est-à-dire une bactérie sur laquelle sera faite la mesure de fluorescence et ses trois aïeules qui ont reçu l'inducteur ; soit P_a le nombre de protéines de type a ($a = O$ ou V pour les gènes *mOrange* ou *egfp*) dans la dernière bactérie de la lignée, à la fin de l'induction, P_{ia}, $i \in [1,4]$, les nombres de protéines produites par chacune des quatre cellules de la lignée, fP^+_{ia}, $i \in [1,3]$, le nombre de protéines qui restent dans la lignée juste après la $i$ème division. Soit $P^+_{ia} = P_{ia} + fP^+_{(i-1)a}$, $i \in [1,3]$, où $fP^+_{0a} = 0$. P^+_{ia} est le nombre de protéines dans la $i$ème cellule à la fin de son cycle, juste avant la $i$ème division (ou à la fin de l'induction).
On peut alors écrire :

$$P_a = P^+_{4a} = P_{4a} + fP^+_{3a}$$

Soit $n_a(t)$ le nombre de copies du gène a à l'instant t, $\alpha_a(i,t)$ le taux de production de protéine a à partir de la copie numéro i du gène a à l'instant t. Le taux de production $\gamma_a(t)$ de protéines peut alors s'écrire :

$$\gamma_a(t) = \sum_{i=1}^{n_a(t)} \alpha_a(i,t)$$

Soit t_i l'âge (le moment du cycle) de la bactérie au début de l'induction. Pour une durée de

3. On pourra consulter par exemple [63], figure 2, où l'on observe une variation typique d'environ 3% du temps de division, dans des conditions moins « favorables » que les cultures en milieu liquide utilisées ici.

4. 20 min est le temps typique de division de bactéries à 37 °C. Dans cette expérience cependant un temps de division de 30 min a été mesuré. Les mêmes expériences ont été répétées à des températures variant de 30 à 39 °C, auxquelles les bactéries croissent à des vitesses différentes. Ainsi, je donnerai aussi les résultats pour $T = 30$ min, c'est-à-dire une induction qui dure deux cycles et des bactéries qui se divisent deux fois pendant l'induction, et pour $T = 1$h (un cycle, une division), temps de division mesuré à 30 °C.

cycle T (ou un taux de croissance $\frac{ln2}{T}$) identique pour chaque bactérie dans la population, t_i suit la distribution $p(t_i) = \frac{2ln2}{T} 2^{-t_i/T}$, $t_i < T$ ([64], volume 2, p1647).

En prenant t_i comme temps initial :

$$P_{1a} = \int_{t_i}^{T} dt\, \gamma_a(t)\ ;\ \ P_{2a} = \int_{T}^{2T} dt\, \gamma_a(t)\ ;\ \ P_{3a} = \int_{2T}^{3T} dt\, \gamma_a(t)\ ;\ \ P_{4a} = \int_{3T}^{3T+t_i} dt\, \gamma_a(t)$$

Remarque : n_a et α_a sont des réalisations particulières, des fonctions *a priori* très bruitées, discontinues (en particulier n_a prend des valeurs entières).

Les crochets $\langle \bullet \rangle$ désignent la moyenne sur l'ensemble des bactéries ($\langle Q \rangle = (1/N_{bact}) \sum_{bact\, i} Q_i$), la barre $\overline{\bullet}$ désigne la moyenne temporelle sur un cycle cellulaire ($\overline{Q} = (1/T) \int_0^T Q(t)dt$).

B.4 Effet des divisions

Le fait que les protéines fluorescentes se répartissent aléatoirement à chaque divsion ajoute des fluctuations aux niveaux de fluorescenec finalement mesurés. À P_{ia}^+ fixé, fP_{ia}^+ suit une loi binomiale de paramètres $(P_{ia}^+, \frac{1}{2})$:

$$p(fP_{ia}^+ | P_{ia}^+) = \frac{1}{2^{P_{ia}^+}} \binom{P_{ia}^+}{fP_{ia}^+}$$

qui vérifie notamment :

$$\begin{aligned} \sum_k k\, p(k|N) &= \frac{N}{2} \\ \sum_k k^2\, p(k|N) &= \frac{N(N+1)}{4} \end{aligned}$$

La répartition des protéines à la division est indépendante de leur production. On peut alors généraliser et simplifier la relation entre probabilités et probabilités conditionnelles :

$$p(\{P_{ja}\}_{j>i}, fP_{ia}^+) = \sum_{P_{ia}^+} p(\{P_{ja}\}_{j>i}, P_{ia}^+).p(fP_{ia}^+ | P_{ia}^+)$$

En notant qu'alors

$$\begin{aligned} p(\{P_{ja}\}_{j>i}, P_{ia}^+) &= \sum_{\substack{P_{ia}, fP_{(i-1)a}^+ \\ P_{ia}+fP_{(i-1)a}^+ = P_{ia}^+}} p(\{P_{ja}\}_{j>i}, P_{ia}, fP_{(i-1)a}^+) \\ &= \sum_{P_{ia}, fP_{(i-1)a}^+} \sum_{P_{(i-1)a}^+} p(\{P_{ja}\}_{j>i-1}, P_{(i-1)a}^+).p(fP_{(i-1)a}^+ | P_{(i-1)a}^+), \\ &\quad P_{ia}+fP_{(i-1)a}^+ = P_{ia}^+ \end{aligned}$$

que la somme sur P_{ia}^+ lève la contrainte $P_{ia} + fP_{(i-1)a} = P_{ia}^+$ et que $P_{1a}^+ = P_{1a}$, on peut écrire la probabilité sur P_a en fonction de la probabilité sur les P_{ia} :

$$\begin{aligned} p(P_a) &= p(P_{4a}^+) \\ &= \sum_{P_{4a}+fP_{3a}^+=P_a} p(P_{4a},P_{3a},P_{2a},P_{1a}).p(fP_{3a}^+|P_{3a}+fP_{2a}^+).p(fP_{2a}^+|P_{2a}+fP_{1a}^+).p(fP_{1a}^+|P_{1a}) \end{aligned}$$

où la somme s'étend sur toutes les valeurs que peuvent prendre les variables sous elle. On peut maintenant écrire les moments de P_a.

$$\begin{aligned} \langle P_a\rangle &= \sum_{P_a} p(P_a)\, P_a \\ &= \sum p(P_{4a},P_{3a},P_{2a},P_{1a}).p(fP_{3a}^+|P_{3a}+fP_{2a}^+).p(fP_{2a}^+|P_{2a}+fP_{1a}^+) \\ &\quad \times p(fP_{1a}^+|P_{1a})\,(P_{4a}+fP_{3a}^+) \end{aligned}$$

En sommant sur les fP_{ia}^+, il vient comme on s'y attend :

$$\boxed{\langle P_a\rangle = \sum_{P_{4a},P_{3a},P_{2a},P_{1a}} p(P_{4a},P_{3a},P_{2a},P_{1a})\left(P_{4a}+\frac{1}{2}P_{3a}+\frac{1}{4}P_{2a}+\frac{1}{8}P_{1a}\right)}$$

Soit $P_a^0 = P_{4a} + \frac{1}{2}P_{3a} + \frac{1}{4}P_{2a} + \frac{1}{8}P_{1a}$. On peut alors écrire [5] :

$$\langle P_a\rangle = \sum_{P_a^0} p(P_a^0)\, P_a^0 = \langle P_a^0\rangle$$

On peut de la même manière évaluer l'autocorrélation du nombre de protéines :

$$\begin{aligned} \langle P_a^2\rangle &= \sum_{P_a} p(P_a)\, P_a^2 \\ &= \sum p(P_{4a},P_{3a},P_{2a},P_{1a}).p(fP_{3a}^+|P_{3a}+fP_{2a}^+).p(fP_{2a}^+|P_{2a}+fP_{1a}^+) \\ &\quad \times p(fP_{1a}^+|P_{1a})(P_{4a}^2+2fP_{3a}^+P_{4a}+(fP_{3a}^+)^2) \\ &= \sum p(P_{4a},P_{3a},P_{2a},P_{1a})\left(P_{4a}^2+\frac{1}{4}P_{3a}^2+\frac{1}{16}P_{2a}^2+\frac{1}{64}P_{1a}^2+P_{3a}P_{4a}\right. \\ &\quad +\frac{1}{2}P_{2a}P_{4a}+\frac{1}{4}P_{2a}P_{3a}+\frac{1}{4}P_{4a}P_{1a}+\frac{1}{8}P_{3a}P_{1a}+\frac{1}{16}P_{1a}P_{2a}+\frac{1}{4}P_{3a}+\frac{3}{16}P_{2a} \\ &\quad \left.+\frac{7}{64}P_{1a}\right) \end{aligned}$$

ce qu'on peut écrire :

$$\boxed{\langle P_a^2\rangle = \langle (P_a^0)^2\rangle + \frac{1}{4}\langle P_{3a}\rangle + \frac{3}{16}\langle P_{2a}\rangle + \frac{7}{64}\langle P_{1a}\rangle}$$

Les divisions contribuent à la variance par des termes « d'ordre P_{ia} », c'est-à-dire qu'elles induisent des fluctuations « d'ordre $\sqrt{P_{ia}}$ » [6].

5. Dans le cas $T = 30$ min : $P_a^0 = P_{3a} + \frac{1}{2}P_{2a} + \frac{1}{4}P_{1a}$. Dans le cas $T = 1$h : $P_a^0 = P_{2a} + \frac{1}{2}P_{1a}$.

6. Dans le cas $T = 30$ min : $\langle P_a^2\rangle = \langle (P_a^0)^2\rangle + \frac{1}{4}\langle P_{2a}\rangle + \frac{3}{16}\langle P_{1a}\rangle$. Dans le cas $T = 1$h : $\langle P_a^2\rangle = \langle (P_a^0)^2\rangle + \frac{1}{4}\langle P_{1a}\rangle$.

Pour $\langle P_a P_b \rangle$, $a \neq b$, P_a et P_b évalués pour la même bactérie, qui fait intervenir $p(P_a, P_b)$, on peut reprendre le même raisonnement que pour $p(P_a)$:

$$\begin{aligned} p(P_a, P_b) &= p(P_{4a}, P_{4b}, fP^+_{3a}, fP^+_{3b}) \\ &= \sum_{P^+_{3a}, P^+_{3b}} p(P_{4a}, P_{4b}, P^+_{3a}, P^+_{3b}).p((fP^+_{3a}, fP^+_{3b})|(P^+_{3a}, P^+_{3b})) \end{aligned}$$

Les protéines EGFP et mOrange se répartissent de façon indépendante lors des divisions. On peut ainsi écrire :

$$p((fP^+_{ia}, fP^+_{jb})|(P^+_{ia}, P^+_{jb})) = p(fP^+_{ia}|P^+_{ia}).p(fP^+_{jb}|P^+_{jb})$$

Le calcul se poursuit de la même manière que pour $\langle P_a \rangle$. On obtient alors :

$$\boxed{\langle P_a P_b \rangle = \sum_{P^0_a, P^0_b} p(P^0_a, P^0_b)\, P^0_a P^0_b = \langle P^0_a P^0_b \rangle}$$

Les divisions n'ajoutent pas de corrélation entre les nombres des deux types de protéines.

B.5 Moyennes

On peut faire l'hypothèse que l'âge de la bactérie au début de l'induction t_i et les fonctions n_a et α_a sont tous indépendants : les probabilités se factorisent : $p[t_i, n_a, \alpha_a] = p(t_i).p[n_a].p[\alpha_a]$. Alors et d'après la section précédente :

$$\begin{aligned} \langle P_a \rangle &= \int dt_i \mathcal{D}[n_a]\mathcal{D}[\alpha_a] p[t_i, n_a, \alpha_a]\, P^0_a[t_i, n_a, \alpha_a] \\ &= \int dt_i\, p(t_i) \left(\frac{1}{8}\int_{t_i}^{T} dt + \frac{1}{4}\int_{T}^{2T} dt + \frac{1}{2}\int_{2T}^{3T} dt + \int_{3T}^{3T+t_i} dt \right) \\ &\quad \times \int \mathcal{D}[n_a] p[n_a] \sum_{i=1}^{n_a(t)} \int \mathcal{D}[\alpha_a] p[\alpha_a] \alpha_a(i,t) \end{aligned}$$

Chaque copie de chromosome ou de pasmide portant la même construction, le taux moyen de production de protéine fluorescente ne dépend pas la copie considérée ; j'ajouterai l'hypothèse qu'en moyenne, ce taux ne dépend pas du temps. Ainsi : $\int \mathcal{D}[\alpha_a] p[\alpha_a] \alpha_a(i,t) = \langle \alpha_a \rangle (i,t)$ est indépendant de i et de t. Avec $\int \mathcal{D}[n_a] p[n_a] \sum_{i=1}^{n_a(t)} = \langle n_a \rangle (t)$, on obtient :

$$\langle P_a \rangle = \langle \alpha_a \rangle \int dt_i\, p(t_i) \left(\frac{1}{8}\int_{t_i}^{T} dt + \frac{1}{4}\int_{T}^{2T} dt + \frac{1}{2}\int_{2T}^{3T} dt + \int_{3T}^{3T+t_i} dt \right) \langle n_a \rangle (t)$$

On peut supposer qu'un régime permanent a été atteint (de nombreuses générations séparent les premières bactéries transformées de celles sur lesquelles ont été faites les mesures) :

$\langle n_a\rangle(t)$ est périodique de période T :

$$\begin{aligned}\langle P_a\rangle &= \langle\alpha_a\rangle \int dt_i\, p(t_i) \left(\frac{1}{8}\int_{t_i}^{T} dt + \frac{3}{4}\int_0^T dt + \int_0^{t_i} dt\right) \langle n_a\rangle(t) \\ &= \frac{7}{8}\langle\alpha_a\rangle \int dt_i\, p(t_i) \left(\int_0^T dt + \int_0^{t_i} dt\right) \langle n_a\rangle(t) \\ &= \frac{7}{8}\, T\langle\alpha_a\rangle \left(\overline{\langle n_a\rangle} + \frac{1}{T}\int dt_i\, p(t_i) \int_0^{t_i} dt\, \langle n_a\rangle(t)\right)\end{aligned}$$

S'il n'y pas de perte de plasmide ou de chromosome, alors n_a est croissante entre deux divisions. $\langle n_a\rangle$ est donc croissante entre 0 et T. On peut de plus supposer qu'en moyenne les plasmides et chromosomes se répartissent de façon symétrique entre deux cellules filles à la division : $\langle n_a\rangle(T) = 2\langle n_a\rangle(0)$.
On peut alors écrire, en posant $\mathcal{R}_a := (\int dt_i\, p(t_i) \int_0^{t_i} dt\, \langle n_a\rangle(t))/\overline{\langle n_a\rangle}$ (voir la section B.7) [7] :

$$\boxed{\langle P_a\rangle \;=\; \frac{7}{8}\, T\langle\alpha_a\rangle(1+\mathcal{R}_a)\overline{\langle n_a\rangle}\ ,\ \text{avec } \mathcal{R}_a \in [0,15; 0,45]}$$

Enfin, les mêmes promoteurs contrôlent l'expression de *egfp* (sur le chromosome) et *mOrange* (plasmide) ; on peut supposer que d'autres effets systématiques n'entrent pas en compte (enroulement de l'ADN, proximité d'autres promoteurs). Enfin, un temps suffisamment long (une nuit) est laissé entre la fin de l'induction et la mesure, si bien que toutes les protéines produites ont atteint leur état natif. On peut alors considérer qu'en moyenne les taux d'expression de *egfp* et de *mOrange* sont identiques : $\langle\alpha_V\rangle = \langle\alpha_O\rangle$. En remarquant enfin que $\overline{\langle n_a\rangle} = \langle\overline{n_a}\rangle$, on peut écrire :

$$\boxed{\frac{\langle\overline{n_O}\rangle}{\langle\overline{n_V}\rangle} \;=\; \frac{1+\mathcal{R}_V}{1+\mathcal{R}_O}\frac{\langle P_O\rangle}{\langle P_V\rangle} \;\in \left[0,79\frac{\langle P_O\rangle}{\langle P_V\rangle}; 1,26\frac{\langle P_O\rangle}{\langle P_V\rangle}\right]}$$

Il faut noter que $\mathcal{R}_a$ dépend de n_a ; plus précisément, on peut voir $\mathcal{R}_a$ comme une indication de la manière dont se réplique le chromosome ou le plasmide au cours d'un cycle. L'incertitude provient ainsi du fait que le chromosome et le plasmide se répliquent au cours de l'induction, sans qu'on sache de quelle façon.

Remarque : on peut ainsi estimer le rapport des moyennes des nombres moyens sur un cycle de copies de plasmide et de chromosme. Des rapports de nombres moyens de plasmides et de chromosome sont aussi mesurés par PCR quantitative (qPCR) ; si ces mesures ne sont pas faites sur des populations de bactéries synchronisées, alors les « nombres moyens » obtenus seront différents des valeurs trouvées ici. Plus précisément, les qPCR font intervenir les

$$\langle n_a\rangle_{\mathrm{q}} \;:=\; \left\langle \int dt_i\, p(t_i)\, n_a(t_i)\right\rangle \;=\; \int dt_i\, p(t_i)\, \langle n_a\rangle(t_i)\ ,\quad a = O, V\ ,$$

7. Dans le cas $T = 30$ min : $\langle P_a\rangle = \frac{3}{4}T\langle\alpha_a\rangle(1+\mathcal{R}_a)\overline{\langle n_a\rangle}$. Dans le cas $T = 1$h : $\langle P_a\rangle = \frac{1}{2}T\langle\alpha_a\rangle(1+\mathcal{R}_a)\overline{\langle n_a\rangle}$.

qui sont différents des $\langle\overline{n_a}\rangle$. Il faut ainsi faire attention en comparant les résultats obtenus par ces deux types de mesures. Un écart a été trouvé entre ces résultats [61], mais il peut venir d'un biais dans la purification des plasmides et du chromosome avant la qPCR. De plus, si l'on peut montrer que $\langle n_a\rangle_{\mathrm{q}}/\langle\overline{n_a}\rangle \in [0,2;0,7]$, il ne semble en revanche pas possible, en l'absence d'information sur la manière dont se répliquent les plasmides et le chromosome, d'estimer *a priori* l'écart entre $\langle n_O\rangle_{\mathrm{q}}/\langle n_V\rangle_{\mathrm{q}}$ et $\langle\overline{n_O}\rangle/\langle\overline{n_V}\rangle$.

B.6 Corrélations

Suivant la même démarche que pour les moyennes, je chercherai à faire ressortir les moments d'ordre 2 de n_a des expressions de ceux de P_a, tout en gardant des termes dépendant de n_a qu'on sait borner.

B.6.1 Corrélations croisées

On s'intéresse ici aux corrélations d'expression de *egfp* (chromosome) et de *mOrange* (plasmide) : $a \neq b$.

$$\begin{aligned}
\langle P_a P_b\rangle &= \int dt_i \mathcal{D}[n_a]\mathcal{D}[n_b]\mathcal{D}[\alpha_a]\mathcal{D}[\alpha_b]p[t_i,n_a,n_b,\alpha_a,\alpha_b] \\
&\quad \times P_a^0[t_i,n_a,\alpha_a]\, P_b^0[t_i,n_b,\alpha_b] \\
&= \int dt_i\, p(t_i) \int \mathcal{D}[n_a]\mathcal{D}[n_b]p[n_a,n_b] \int \mathcal{D}[\alpha_a]\mathcal{D}[\alpha_b]p[\alpha_a,\alpha_b] \\
&\quad \times \left(\frac{1}{8}\int_{t_i}^{T} dt_a + \frac{1}{4}\int_{T}^{2T} dt_a + \frac{1}{2}\int_{2T}^{3T} dt_a + \int_{3T}^{3T+t_i} dt_a\right) \sum_{i_a=1}^{n_a(t_a)} \alpha_a(i_a,t_a) \\
&\quad \times \left(\frac{1}{8}\int_{t_i}^{T} dt_b + \frac{1}{4}\int_{T}^{2T} dt_b + \frac{1}{2}\int_{2T}^{3T} dt_b + \int_{3T}^{3T+t_i} dt_b\right) \sum_{i_b=1}^{n_b(t_b)} \alpha_b(i_b,t_b)
\end{aligned}$$

avec comme précédemment l'hypothèse que l'âge de la bactérie, les nombres de copies et les taux d'expression sont indépendants : $p[t_i,n_a,n_b,\alpha_a,\alpha_b] = p(t_i).p[n_a,n_b].p[\alpha_a,\alpha_b]$.

On supposera de plus que les corrélations d'expression ne dépendent en moyenne, pour des types de gènes différents, ni des copies considérés ni du temps :

$$\int \mathcal{D}[\alpha_a]\mathcal{D}[\alpha_b]p[\alpha_a,\alpha_b]\alpha_a(i_a,t_a)\alpha_b(i_b,t_b) = \langle\alpha_a\alpha_b\rangle(i_a,t_a;i_b,t_b) = \langle\alpha_O\alpha_V\rangle,$$

pour $a \neq b$. Enfin, on peut admettre que les corrélations des nombres de copies de chromosome et de plasmide à deux temps ne dépendent en moyenne que des moments du cycle

considéré : $\langle n_a n_b\rangle(t_a;t_b)$ est périodique de prériode T par rapport à t_a et t_b. Alors :

$$\begin{aligned}\langle P_a P_b\rangle &= \frac{49}{64}\langle\alpha_O\alpha_V\rangle\int dt_i\, p(t_i)\left(\int_0^T dt_a+\int_0^{t_i} dt_a\right)\left(\int_0^T dt_b+\int_0^{t_i} dt_b\right)\langle n_a n_b\rangle(t_a;t_b)\\ &= \frac{49}{64}\langle\alpha_O\alpha_V\rangle\left(T^2\overline{\langle n_a\rangle}\;\overline{\langle n_b\rangle}+T\int dt_i\, p(t_i)\int_0^{t_i} dt(\overline{\langle n_a\rangle}n_b\rangle(;t)+\langle n_a\overline{n_b\rangle}(t;))\right.\\ &\quad\left.+\int dt_i\, p(t_i)\int_0^{t_i} dt_a\int_0^{t_i} dt_b\langle n_a n_b\rangle(t_a;t_b)\right)\\ &= \frac{49}{64}\langle\alpha_O\alpha_V\rangle T^2(1+\mathcal{R}_a+\mathcal{R}_b+\mathcal{S}_{ab})\overline{\langle n_a\rangle}\;\overline{\langle n_b\rangle}\end{aligned}$$

où $\mathcal{S}_{ab} := (\int dt_i\, p(t_i)\int_0^{t_i} dt_a\int_0^{t_i} dt_b\langle n_a n_b\rangle(t_a;t_b))/\overline{\langle n_a\rangle}\;\overline{\langle n_b\rangle}$.
On peut montrer que $\mathcal{S}_{ab}\in[0;0,45]$ (voir en section B.7). En remarquant [8] que $\overline{\langle n_a\rangle}\;\overline{\langle n_b\rangle} = \langle\overline{n_a}\;\overline{n_b}\rangle$, on peut écrire [9] :

$$\boxed{\langle P_O P_V\rangle = \frac{49}{64}T^2\langle\alpha_O\alpha_V\rangle(1+\mathcal{R}_O+\mathcal{R}_V+\mathcal{S}_{OV})\langle\overline{n_O}\;\overline{n_V}\rangle}$$

avec $\mathcal{R}_O$, $\mathcal{R}_V\in[0,15;0,45]$ et $\mathcal{S}_{OV}\in[0;0,45]$. $\mathcal{S}_{ab}$ mesure à la fois la manière dont le plasmides et le chromosome se répliquent et se répartissent à la division.

B.6.2 Autocorrélations

D'après la section B.4 :

$$\langle P_a^2\rangle = \langle(P_a^0)^2\rangle+\frac{1}{4}\langle P_{3a}\rangle+\frac{3}{16}\langle P_{2a}\rangle+\frac{7}{64}\langle P_{1a}\rangle$$

On peut évaluer la contribution de la répartition des protéines lors des divisions :

$$\begin{aligned}\langle P_{3a}\rangle = \langle P_{2a}\rangle &= T\langle\alpha_a\rangle\overline{\langle n_a\rangle} = \frac{8}{7}\frac{1}{1+\mathcal{R}_a}\langle P_a\rangle\\ \langle P_{1a}\rangle &= \langle\alpha_a\rangle\int dt_i\, p(t_i)\int_{t_i}^T dt\,\langle n_a\rangle(t)\\ &= T\langle\alpha_a\rangle\left(\overline{\langle n_a\rangle}-\frac{1}{T}\int dt_i\, p(t_i)\int_0^{t_i} dt\,\langle n_a\rangle(t)\right)\\ &= T\langle\alpha_a\rangle(1-\mathcal{R}_a)\overline{\langle n_a\rangle}\\ &= \frac{8}{7}\left(\frac{1-\mathcal{R}_a}{1+\mathcal{R}_a}\right)\langle P_a\rangle\end{aligned}$$

8.
$$\begin{aligned}\overline{\langle n_a\rangle}\;\overline{\langle n_b\rangle} &= \frac{1}{T^2}\int_0^T dt_1\int_0^T dt_2\int\mathcal{D}[n_a]p[n_a]n_a(t_1)n_b(t_2)\\ &= \int\mathcal{D}[n_a]p[n_a]\frac{1}{T}\int_0^T dt_1\, n_a(t_1)\frac{1}{T}\int_0^T dt_2\, n_b(t_1)\\ &= \langle\overline{n_a}\;\overline{n_b}\rangle\end{aligned}$$

9. Dans le cas $T=30$min : $\langle P_O P_V\rangle = \frac{9}{16}T^2\langle\alpha_O\alpha_V\rangle(1+\mathcal{R}_O+\mathcal{R}_V+\mathcal{S}_{OV})\langle\overline{n_O}\;\overline{n_V}\rangle$. Dans le cas $T=1$h : $\langle P_O P_V\rangle = \frac{1}{4}T^2\langle\alpha_O\alpha_V\rangle(1+\mathcal{R}_O+\mathcal{R}_V+\mathcal{S}_{OV})\langle\overline{n_O}\;\overline{n_V}\rangle$.

Donc [10] :

$$\boxed{\langle P_a^2\rangle \;=\; \langle (P_a^0)^2\rangle + \frac{7}{64}T\langle\alpha_a\rangle(5-\mathcal{R}_a)\overline{\langle n_a\rangle} \;=\; \langle (P_a^0)^2\rangle + \frac{1}{8}\left(\frac{5-\mathcal{R}_a}{1+\mathcal{R}_a}\right)\langle P_a\rangle}$$

Il reste à évaluer $\langle (P_a^0)^2\rangle$.

On peut faire l'hypothèse que, à un instant donné, les expressions de deux copies différentes d'un gène a corrèlent comme les expressions de deux gènes de types différents : les corrélations entre deux plasmides (deux chromosomes) sont les mêmes qu'entre un plasmide et un chromosome.

On peut faire deux hypothèses extrêmes sur les corrélations d'expression à différents temps : (A) il n'y a de mémoire que sur un temps τ très court, ou (B) une copie corrèle avec elle-même (ou ses ancêtres) de la même manière sur toute la durée d'induction. La première est la plus réaliste, le temps de corrélation d'une copie avec elle-même étant vraisemblablement de l'ordre de la durée de vie d'un ARN (des corrélations beaucoup plus courtes peuvent aussi exister, provenant par exemple de la présence d'ARN-polymérases ou de facteurs de transcriptions autour du promoteur, qui peuvent s'y fixer plusieurs fois avant de diffuser). La deuxième hypothèse, beaucoup moins réaliste, correspondrait par exemple à des mutations ponctuelles distinguant des copies de promoteur au début de l'induction et qui seraient transmises à leurs descendants ; elle servira surtout à estimer la sensibilité des résultats à l'hypothèse faite sur les temps de corrélation.

Une situation intermédiaire entre ces deux hypothèses pourrait correspondre à différents enroulements des plamsides à leur création : le temps de corrélation serait alors de l'ordre du temps de réplication d'un plasmide, c'est-à-dire de la durée d'un cycle cellulaire.

Comme pour les corrélations croisées, je ferai l'hypothèse que les corrélations d'expression, qu'il s'agisse d'une copie avec elle-même ou d'une copie avec une autre, ne dépendent en moyenne ni des copies, ni du temps, ni du type de gène. Je noterai ainsi $\langle\alpha^2\rangle$ l'autocorrélation d'expression d'une copie de gène et admettrai que deux copies différentes d'un même gène corrèlent comme deux copies de gènes différents : $\langle\alpha_a\alpha_a\rangle(i_1,t_1;i_2,t_2) = \langle\alpha_O\alpha_V\rangle,\ \forall t_1,t_2,i_1\neq i_2$.

Hypothèse (A) : Cette hypothèse n'a de sens que si τ est très inférieur au temps typique de réplication du plasmide ou du chromosome, ce qui pour τ de l'ordre de la minute est bien le cas (chaque copie de gène est en moyenne répliquée une fois au cours du cycle). Au-delà du temps τ, la corrélation d'expression d'une copie de gène avec elle-même est du même type qu'entre deux copies différentes. On peut traduire cette hypothèse ainsi :

$$\langle\alpha_a\alpha_a\rangle(i_1,t_1;i_2,t_2) = \langle\alpha^2\rangle\tau\delta(t_2-t_1)\delta_{i_1i_2} + \langle\alpha_O\alpha_V\rangle(1-\tau\delta(t_2-t_1)\delta_{i_1i_2})$$

10. Dans le cas $T = 30$ min : $\langle P_a^2\rangle = \langle (P_a^0)^2\rangle + \frac{1}{16}T\langle\alpha_a\rangle(7-3\mathcal{R}_a)\overline{\langle n_a\rangle} = \langle (P_a^0)^2\rangle + \frac{1}{12}\left(\frac{7-3\mathcal{R}_a}{1+\mathcal{R}_a}\right)\langle P_a\rangle$.
Dans le cas $T = 1$h : $\langle P_a^2\rangle = \langle (P_a^0)^2\rangle + \frac{1}{4}T\langle\alpha_a\rangle(1-\mathcal{R}_a)\overline{\langle n_a\rangle} = \langle (P_a^0)^2\rangle + \frac{1}{2}\left(\frac{1-\mathcal{R}_a}{1+\mathcal{R}_a}\right)\langle P_a\rangle$.

On obtient alors, en faisant comme précédemment l'hypothèse de périodicité de $\langle n_a n_a\rangle$:

$$\begin{aligned}\langle (P_a^0)^2\rangle_{\rm A} &= \frac{49}{64}\langle\alpha_O\alpha_V\rangle\int dt_i\, p(t_i)\left\{\left(\int_0^T dt_1+\int_0^{t_i} dt_1\right)\left(\int_0^T dt_2+\int_0^{t_i} dt_2\right)\langle n_a n_a\rangle(t_1;t_2)\right.\\ &\quad +(\langle\alpha^2\rangle-\langle\alpha_O\alpha_V\rangle)\tau\left(\frac{1}{8}\int_{t_i}^T dt_1+\frac{1}{4}\int_T^{2T} dt_1+\frac{1}{2}\int_{2T}^{3T} dt_1+\int_{3T}^{3T+t_i} dt_1\right)\\ &\quad \left.\times\left(\frac{1}{8}\int_{t_i}^T dt_2+\frac{1}{4}\int_T^{2T} dt_2+\frac{1}{2}\int_{2T}^{3T} dt_2+\int_{3T}^{3T+t_i} dt_2\right)\delta(t_2-t_1)\langle n_a\rangle(t_1)\right\}\\ &= \frac{49}{64}T^2\langle\alpha_O\alpha_V\rangle(1+2\mathcal{R}_a+\mathcal{S}_{aa})\overline{\langle n_a}\,\overline{n_a\rangle}\\ &\quad +\frac{21}{64}\tau(\langle\alpha^2\rangle-\langle\alpha_O\alpha_V\rangle)\int dt_i\, p(t_i)\left(\int_0^T dt_1+3\int_0^{t_i} dt_1\right)\langle n_a\rangle(t_1)\end{aligned}$$

Soit [11] :

$$\boxed{\langle (P_a^0)^2\rangle_{\rm A} = \frac{49}{64}T^2\langle\alpha_O\alpha_V\rangle(1+2\mathcal{R}_a+\mathcal{S}_{aa})\langle(\overline{n_a})^2\rangle+\frac{21}{64}\tau T(\langle\alpha^2\rangle-\langle\alpha_O\alpha_V\rangle)(1+3\mathcal{R}_a)\langle\overline{n_a}\rangle}$$

Hypothèse (B) : Considérons deux instants t_1 et t_2, $t_2>t_1$, de l'induction. Soit une copie du gène a à t_2 ; parmi les copies à t_1, une seule est son ancêtre, les $n_a(t_1)-1$ autres copies ne lui étant pas liées. Or cela est vrai pour les $n_a(t_2)$ copies à t_2. Ainsi, dans cette hypothèse :

$$\begin{aligned}\sum_{i_1=1}^{n_a(t_1)}\sum_{i_2=1}^{n_a(t_2)}\langle\alpha_a\alpha_a\rangle(i_1,t_1;i_2,t_2) &= \begin{cases} n_a(t_2)(\langle\alpha^2\rangle+(n_a(t_1)-1)\langle\alpha_O\alpha_V\rangle\ ,\ \text{si } t_2\geq t_1\\ n_a(t_1)(\langle\alpha^2\rangle+(n_a(t_2)-1)\langle\alpha_O\alpha_V\rangle\ ,\ \text{si } t_2\leq t_1\end{cases}\\ &= \langle\alpha_O\alpha_V\rangle n_a(t_1)n_a(t_2)\\ &\quad +(\langle\alpha^2\rangle-\langle\alpha_O\alpha_V\rangle)(n_a(t_1)\Theta(t_1-t_2)+n_a(t_2)\Theta(t_2-t_1))\end{aligned}$$

Alors :

$$\begin{aligned}\langle (P_a^0)^2\rangle_{\rm B} &= \frac{49}{64}T^2\langle\alpha_O\alpha_V\rangle(1+2\mathcal{R}_a+\mathcal{S}_{aa})\overline{\langle n_a}\,\overline{n_a\rangle}\\ &\quad +(\langle\alpha^2\rangle-\langle\alpha_O\alpha_V\rangle)\int dt_i\, p(t_i)\left(\frac{1}{8}\int_{t_i}^T dt_1+\frac{1}{4}\int_T^{2T} dt_1+\frac{1}{2}\int_{2T}^{3T} dt_1\right.\\ &\quad \left.+\int_{3T}^{3T+t_i} dt_1\right)\left(\frac{1}{8}\int_{t_i}^T dt_2+\frac{1}{4}\int_T^{2T} dt_2+\frac{1}{2}\int_{2T}^{3T} dt_2+\int_{3T}^{3T+t_i} dt_2\right)\\ &\quad \times(\langle n_a\rangle(t_1)\Theta(t_1-t_2)+\langle n_a\rangle(t_2)\Theta(t_2-t_1))\\ &= \frac{49}{64}T^2\langle\alpha_O\alpha_V\rangle(1+2\mathcal{R}_a+\mathcal{S}_{aa})\overline{\langle n_a}\,\overline{n_a\rangle}+\frac{1}{32}(\langle\alpha^2\rangle-\langle\alpha_O\alpha_V\rangle)\int dt_i\, p(t_i)\\ &\quad \times\left(35\int_0^T dt_1\, t_1+119\int_0^{t_i} dt_1\, t_1+(18T-7t_i)\int_0^T dt_1\right.\\ &\quad \left.+(88T-7t_i)\int_0^{t_i} dt_1\right)\langle n_a\rangle(t_1)\end{aligned}$$

11. Dans le cas $T=30$min : $\langle (P_a^0)^2\rangle_{\rm A}=\frac{9}{16}T^2\langle\alpha_O\alpha_V\rangle(1+2\mathcal{R}_a+\mathcal{S}_{aa})\langle(\overline{n_a})^2\rangle+\frac{5}{16}\tau T(\langle\alpha^2\rangle-\langle\alpha_O\alpha_V\rangle)(1+3\mathcal{R}_a)\langle\overline{n_a}\rangle$. Dans le cas $T=1$h : $\langle (P_a^0)^2\rangle_{\rm A}=\frac{1}{4}T^2\langle\alpha_O\alpha_V\rangle(1+2\mathcal{R}_a+\mathcal{S}_{aa})\langle(\overline{n_a})^2\rangle+\frac{1}{4}\tau T(\langle\alpha^2\rangle-\langle\alpha_O\alpha_V\rangle)(1+3\mathcal{R}_a)\langle\overline{n_a}\rangle$.

où on a utilisé la la symétrie $t_1 \leftrightarrow t_2$, la périodicité de $\langle n_a \rangle$ et fait les changements de variable $t_1 \to t_1 - T/2T/3T$. Avec le même type d'analyse que précédemment, définissant une grandeur $\mathcal{T}_a$ dont on montre qu'elle appartient à l'intervalle $[0,48;7,8]$ (voir en section B.7), on trouve [12] :

$$\boxed{\langle (P_a^0)^2 \rangle_{\mathrm{B}} = \frac{49}{64} T^2 \langle \alpha_O \alpha_V \rangle (1 + 2\mathcal{R}_a + \mathcal{S}_{aa}) \langle (\overline{n_a})^2 \rangle + \frac{9}{16} T^2 (\langle \alpha^2 \rangle - \langle \alpha_O \alpha_V \rangle)(1 + \mathcal{T}_a) \langle \overline{n_a} \rangle}$$

Les expressions de $\langle (P_a^0)^2 \rangle$ dans chaque hypothèse diffèrent par le coefficient du terme en $\langle \overline{n_a} \rangle$. Le rapport de ces deux coefficients vaut :

$$\frac{7}{12} \left(\frac{1 + 3\mathcal{R}_a}{1 + \mathcal{T}_a} \right) \frac{\tau}{T} \in \left[0, 3\frac{\tau}{T}; \frac{\tau}{T} \right]$$

en considérant que $\mathcal{R}_a$ et $\mathcal{T}_a$ sont indépendants. La taille de l'intervalle auquel appartient $\mathcal{T}_a$ est déjà surestimé : la taille de l'intervalle auquel appartient le rapport précédent est très largement surestimé... [13]

B.7 Simplification des expressions des moments de P_a

B.7.1 Estimation de $\mathcal{R}_a$

Soit :

$$\begin{aligned} N_a(t) &:= \frac{1}{T} \int_0^t dt \, \langle n_a \rangle(t) \\ R_{a,1} &:= \frac{\overline{N_a}}{\overline{\langle n_a \rangle}} \\ R_{a,2} &:= \frac{1}{T^2} \frac{1}{\overline{\langle n_a \rangle}} \int_0^T dt_i \, t_i N_a(t_i) \\ \mathcal{R}_a &:= \frac{1}{\overline{\langle n_a \rangle}} \int dt_i \, p(t_i) \, N_a(t_i) \end{aligned}$$

de telle sorte que $\langle P_a \rangle = \frac{7}{8} T \langle \alpha_a \rangle (1 + \mathcal{R}_a) \overline{\langle n_a \rangle}$. Pour simplifier le calcul, j'utiliserai le développement à l'ordre 1 en t_i/T de $p(t_i)$:

$$\mathcal{R}_a \approx 2ln2(R_{a,1} - ln2 \, R_{a,2})$$

On peut admettre que $\langle n_a \rangle$ est croissante entre 0 et T et $\langle n_a \rangle(T) = 2\langle n_a \rangle(0)$. N_a est donc convexe : en particulier, $\forall t, N_a(t) \leq t(N_a(T) - N_a(0))/T - N_a(0) = (t/T)N_a(T) =$

12. Dans le cas $T = 30$ min : $\langle (P_a^0)^2 \rangle_{\mathrm{B}} = \frac{9}{16} T^2 \langle \alpha_O \alpha_V \rangle (1 + 2\mathcal{R}_a + \mathcal{S}_{aa}) \langle (\overline{n_a})^2 \rangle + \frac{1}{8} T^2 (\langle \alpha^2 \rangle - \langle \alpha_O \alpha_V \rangle)(1 + \mathcal{T}_a) \langle \overline{n_a} \rangle$, et la définition de $\mathcal{T}_a$ diffère (voir en section B.7). Dans le cas $T = 1$h : $\langle (P_a^0)^2 \rangle_{\mathrm{B}} = \frac{1}{4} T^2 \langle \alpha_O \alpha_V \rangle (1 + 2\mathcal{R}_a + \mathcal{S}_{aa}) \langle (\overline{n_a})^2 \rangle + \frac{1}{4} T^2 (\langle \alpha^2 \rangle - \langle \alpha_O \alpha_V \rangle)(1 + \mathcal{T}_a) \langle \overline{n_a} \rangle$, et la définition de $\mathcal{T}_a$ diffère (voir en section B.7).

13. Pour $T = 30$ min, ce rapport est $\frac{5}{2} \left(\frac{1 + 3\mathcal{R}_a}{1 + \mathcal{T}_a} \right) \frac{\tau}{T} \in \left[0, 3\frac{\tau}{T}; 6\frac{\tau}{T} \right]$. Pour $T = 1$h, l'estimation de $\mathcal{T}_a$ obtenue ne permet pas de majorer ce rapport.

$(t/T)\overline{\langle n_a\rangle}$. D'où, en intégrant sur t :

$$R_{a,1} \leq \frac{1}{2}$$

Il s'ensuit également que $\forall t, N_a(t) \geq (t/T)\langle n_a\rangle(0)$, soit, après intégration sur t et division par T : $\overline{N_a} \geq \langle n_a\rangle(0)/2$, et après division par $\overline{\langle n_a\rangle}$:

$$R_{a,1} \geq \frac{\langle n_a\rangle(0)}{2\overline{\langle n_a\rangle}}$$

Comme $\langle n_a\rangle$ est strictement positive, on peut faire le changement de variable $t \to u = N_a(t)$:

$$\overline{N_a} = \int_0^{N_a(T)} du\, \frac{u}{\langle n_a\rangle(N_a^{-1}(u))}.$$

N_a est strictement croissante, donc N_a^{-1} aussi ; de plus, $1/\langle n_a\rangle$ est décroissante ; donc $1/\left(\langle n_a\rangle \circ N_a^{-1}\right)$ est décroissante. D'où :

$$\overline{N_a} \;\geq\; \frac{1}{\langle n_a\rangle \circ N_a^{-1}(N_a(T))}\int_0^{N_a(t)} du\, u \;=\; \frac{1}{\langle n_a\rangle(T)}\frac{1}{2}N_a(T)^2 \;=\; \frac{\overline{\langle n_a\rangle}^2}{2\langle n_a\rangle(T)}.$$

Soit :

$$R_{a,1} \geq \frac{\overline{\langle n_a\rangle}}{2\langle n_a\rangle(T)}$$

En multipliant les deux minorants de $R_{a,1}$, en prenant la racine carrée et en utilisant l'hypothèse $\langle n_a\rangle(T) = 2\langle n_a\rangle(0)$, il vient : $R_{a,1} \geq 1/2\sqrt{2}$. Soit :

$$R_{a,1} \in \left[\frac{\sqrt{2}}{4}, \frac{1}{2}\right]$$

Il existe $t_i^0 \in [0,T]$ tel que $\int_0^T dt_i\, t_i\langle N_a\rangle(t_i) = t_i^0 \int_0^T dt_i\, \langle N_a\rangle(t_i)$. Comme N_a est croissante, alors $\int_0^T dt_i\, (t_i - T/2)N_a(t_i) \geq 0$. En effet :

$$\begin{aligned}
\int_0^T dt_i \left(t_i - \frac{T}{2}\right) N_a(t_i) &= \left(\int_0^{\frac{T}{2}} dt_i + \int_{\frac{T}{2}}^T dt_i\right)\left(t_i - \frac{T}{2}\right) N_a(t_i) \\
&= N_a(t_1)\int_0^{\frac{T}{2}} dt_i \left(t_i - \frac{T}{2}\right) + N_a(t_2)\int_{\frac{T}{2}}^T dt_i \left(t_i - \frac{T}{2}\right), \\
&\quad \text{où } t_1 \in \left[0; \frac{T}{2}\right] \text{ et } t_2 \in \left[\frac{T}{2}; T\right] \\
&= \frac{T^2}{8}(N_a(t_2) - N_a(t_1)) \;>\; 0
\end{aligned}$$

Donc $t_i^0 \geq T/2$. D'où :

$$R_{a,2} = \frac{t_i^0}{T}R_{a,1}\ ,\ \text{avec } t_i^0 \in \left[\frac{T}{2}, T\right]$$

Il s'ensuit $\mathcal{R}_a \approx 2ln2(1 - t_i^0(ln2/T))R_{a,1}$, soit :

$$\boxed{\mathcal{R}_a \in [0,15; 0,45]}$$

B.7.2 Estimation de $\mathcal{S}_{ab}$

Soit :

$$\begin{aligned} G_{ab}(t) &:= \frac{1}{T^2}\int_0^t dt_1 \int_0^t dt_2\ \langle n_a n_b\rangle(t_1;t_2) \\ \mathcal{S}_{ab} &:= \frac{1}{\overline{\langle n_a\rangle}\ \overline{n_b\rangle}} \int dt_i\, p(t_i)\ G_{ab}(t_i) \end{aligned}$$

En remarquant que $G_{ab}(T) = \overline{\langle n_a}\ \overline{n_b\rangle}$, il vient, de la même manière que dans la section précédente (G_{ab} est croissante) :

$$\mathcal{S}_{ab} \approx 2ln2\left(1 - t^0\frac{ln2}{T}\right)\frac{\overline{G_{ab}}}{G_{ab}(T)}\ ,\ \text{avec } t^0 \in \left[\frac{T}{2};T\right].$$

La dérivée de G_{ab} s'écrit [14] :

$$G'_{ab}(t) = \frac{1}{T^2}\int_0^t du\,(\langle n_a n_b\rangle(t;u) + \langle n_a n_b\rangle(u;t))$$

On voit que G'_{ab} est croissante, donc G_{ab} est convexe. Comme précédemment, on peut en déduire : $\frac{\overline{G_{ab}}}{G_{ab}(T)} \leq \frac{1}{2}$. Par contre $G'_{ab}(0) = 0$: il semble difficile d'obtenir mieux que $\frac{\overline{G_{ab}}}{G_{ab}(T)} \geq 0$ dans le cas général.

D'après l'expression précédente de $\mathcal{S}_{ab}$, tenant compte de ces bornes, on obtient :

$$\boxed{\mathcal{S}_{ab} \in [0;0,45]}$$

Remarque : on peut calculer la dérivée seconde de G_{ab} (expression vérifiée de la même manière que celle de G'_{ab}) :

$$G''_{ab}(t) = \frac{2}{T^2}\langle n_a n_b\rangle(t;t) + \frac{1}{T^2}\int_0^t du\,\left(\frac{d}{dt_a}\langle n_a n_b\rangle(t;u) + \frac{d}{dt_b}\langle n_a n_b\rangle(u;t)\right),$$

positive mais pas manifestement monotone.

B.7.3 Estimation de $\mathcal{T}_a$

Soit :

$$\begin{aligned} \mathcal{T}_a &:= \frac{1}{\overline{\langle n_a\rangle}}\frac{1}{18T^2}\int dt_i p(t_i)\left(35\int_0^T dt\,t + 119\int_0^{t_i} dt\,t - 7t_i\int_0^T dt + 88T\int_0^{t_i} dt\right. \\ &\quad \left. -7t_i\int_0^{t_i} dt\right)\langle n_a\rangle(t) \\ \overline{t\langle n_a\rangle} &:= \frac{1}{T}\int_0^T dt\,t\langle n_a\rangle(t) = [t\langle N_a\rangle(t)]_0^T - \int_0^T dt\,\langle N_a\rangle(t) = T(1-R_{a,1})\overline{\langle n_a\rangle} \\ \widetilde{\mathcal{R}_a} &:= \frac{1}{\overline{t\langle n_a\rangle}}\frac{1}{T}\int dt_i p(t_i)\int_0^{t_i} dt\,t\langle n_a\rangle(t) \end{aligned}$$

14. On peut montrer que $\int_0^t G'_{ab}(t')dt' = G(t)$ en symétrisant les deux termes et en remarquant qu'alors le terme de droite fait apparaître l'intégrale sur le carré de côté $\{(0;0),(t;0),(t;t),(0;t)\}$ et le terme de gauche la somme des intégrales sur les triangles $\{(0;0),(t;0),(t;t)\}$ et $\{(0;0),(0;t),(t;t)\}$ de la même fonction.

En procédant de la même manière que dans les sections précédentes, on peut simplifier l'expression de $\mathcal{T}_a$:

$$\begin{aligned}\mathcal{T}_a &\approx \frac{1}{\overline{\langle n_a\rangle}}\frac{1}{18}\left(35\frac{1}{T}\overline{t\langle n_a\rangle} - 7(2ln2-(ln2)^2)\overline{\langle n_a\rangle} + 88\mathcal{R}_a\overline{\langle n_a\rangle} + \frac{119}{T}\widetilde{\mathcal{R}_a}\overline{t\langle n_a\rangle}\right.\\ &\left. -7\frac{2ln2}{T^2}\int_0^T dt_i\,\left(t_i - t_i^2\frac{ln2}{T}\right)\langle N_a\rangle(t_i)\right)\end{aligned}$$

Il existe $t_i^1 \in [\frac{T}{2};T]$ tel que :

$$\begin{aligned}\int_0^T dt_i\, t_i^2\langle N_a\rangle(t_i) &= t_i^1\int_0^T dt_i\, t_i\langle N_a\rangle(t_i)\\ &= t_i^1\, t_i^0\, R_{a,1}T\overline{\langle n_a\rangle}\end{aligned}$$

où t_i^0 a été défini lors de l'estimation de $\mathcal{R}_a$.
Enfin, $\widetilde{\mathcal{R}_a} \in [0;0,45]$.

$$\begin{aligned}\mathcal{T}_a &\approx \frac{1}{18}\left(35(1-R_{a,1}) + 119\widetilde{\mathcal{R}_a}(1-R_{a,1}) - 7(2ln2-(ln2)^2) + 88\times 2ln2\left(1-t_i^0\frac{ln2}{T}\right)R_{a,1}\right.\\ &\left. -\frac{14ln2}{T}\left(t_i^0 - t_i^1\, t_i^0\frac{ln2}{T}\right)R_{a,1}\right)\\ &= \frac{7}{18}(5-2ln2+(ln2)^2) + \frac{119}{18}\widetilde{\mathcal{R}_a} + \frac{R_{a,1}}{18}\left(176ln2 - 35 - 119\widetilde{\mathcal{R}_a}\right.\\ &\left. + \left(14(ln2)^2\frac{t_i^1}{T} - 14ln2 - 176(ln2)^2\right)\frac{t_i^0}{T}\right)\end{aligned}$$

Soit, en considérant tous les paramètres comme indépendants (ce qui surestime la taille de l'intervalle) [15] :

$$\boxed{\mathcal{T}_a \in [0,47;7,8]}$$

B.7.4 Avec des fonctions tests

Pour se faire une idée de la qualité des estimations précédentes et fixer des intervalles minimaux, on peut calculer $\mathcal{R}_a$, $\mathcal{S}_{ab}$ et $\mathcal{T}_a$ en postulant différentes formes pour $\langle n_a\rangle$ et $\langle n_a n_b\rangle$.

15. Dans le cas $T = 30$ min :

$$\mathcal{T}_a \approx \frac{1}{\overline{\langle n_a\rangle}}\frac{1}{2T^2}\int dt_i p(t_i)\left(7\int_0^T dt\, t + 27\int_0^{t_i} dt\, t - 3t_i\int_0^T dt + 16T\int_0^{t_i} dt - 3t_i\int_0^{t_i} dt\right)\langle n_a\rangle(t)$$

ce qui conduit à :

$$\mathcal{T}_a \in [0;9,9]$$

Dans le cas $T = 1$h :

$$1+\mathcal{T}_a \approx \frac{1}{\overline{\langle n_a\rangle}}\frac{1}{T^2}\int dt_i p(t_i)\left(\int_0^T dt\, t + 5\int_0^{t_i} dt\, t - t_i\int_0^T dt + 4T\int_0^{t_i} dt - t_i\int_0^{t_i} dt\right)\langle n_a\rangle(t)$$

ce qui conduit à :

$$1+\mathcal{T}_a \in [0;4,4]$$

Les changements de variables $t \to t/T$ et $t_i \to t_i/T$, et la normalisation $\langle n_a \rangle \to \langle n_a \rangle / \langle n_a \rangle(0)$ laissent $\mathcal{R}_a$ et $\mathcal{T}_a$ inchangés. On peut donc considérer des fonctions croissantes définies sur $[0,1]$ et variant de 1 à 2.

Le calcul a été fait pour des fonctions marche, sigmoïde, exponentielle, logarithme, sinus et des monômes de degré quelconque. Chaque type de fonction est défini par un ou deux paramètres (par exemple le temps auquel la fonction marche passe de 1 à 2 ou la constante de temps de l'exponentielle) : six ou quatre valeurs ont été attribuées à chacun, selon que la fonction dépend d'un ou deux paramètres.

Pour $\mathcal{S}_{ab}$, j'ai considéré des produits de deux fonctions quelconques parmi les précédentes. Cela implique cependant notamment $\langle n_a n_b \rangle(T) = 4\langle n_a n_b \rangle(0)$, ce qui n'est en général pas vrai.

Toutes les intégrales ont été estimées par la méthode des rectangles avec des pas d'intégration de taille $1/1000$ ($1/100$ pour $\mathcal{S}_{ab}$) et l'expression exacte de $p(t_i)$.

On trouve alors [16] :

$$\mathcal{R}_a^{\text{test}} \in [0,36; 0,44]; \quad \mathcal{S}_{ab}^{\text{test}} \in [0,20; 0,28]; \quad \mathcal{T}_a^{\text{test}} \in [3,5; 3,8]$$

Si les bornes trouvées pour $\mathcal{R}_a$ semblent assez bonnes, ces estimations suggèrent que la taille des intervalles auquels appartiennent $\mathcal{S}_{ab}$ et surtout $\mathcal{T}_a$ pourrait être réduite.

B.8 Discussion

B.8.1 Synthèse

En notant K_i les constantes (indépendantes du type de gène), on peut écrire, dans l'hypothèse (A) :

$$\begin{aligned}
\langle P_O \rangle &= K_1(1+\mathcal{R}_O)\langle \overline{n_O} \rangle \\
\langle P_V \rangle &= K_1(1+\mathcal{R}_V)\langle \overline{n_V} \rangle \\
\langle P_O P_V \rangle &= K_2(1+\mathcal{R}_O+\mathcal{R}_V+\mathcal{S}_{OV})\langle \overline{n_O}\ \overline{n_V} \rangle \\
\langle P_O^2 \rangle_{\text{A}} &= \frac{1}{8}K_1(5-\mathcal{R}_O)\langle \overline{n_O} \rangle + K_3(1+3\mathcal{R}_O)\langle \overline{n_O} \rangle + K_2(1+2\mathcal{R}_O+\mathcal{S}_{OO})\langle (\overline{n_O})^2 \rangle \\
\langle P_V^2 \rangle_{\text{A}} &= \frac{1}{8}K_1(5-\mathcal{R}_V)\langle \overline{n_V} \rangle + K_3(1+3\mathcal{R}_V)\langle \overline{n_V} \rangle + K_2(1+2\mathcal{R}_V+\mathcal{S}_{VV})\langle (\overline{n_V})^2 \rangle
\end{aligned}$$

où $K_1 := \frac{7}{8}T\langle \alpha \rangle$, $K_2 := \frac{49}{64}T^2\langle \alpha_O \alpha_V \rangle$, $K_3 := \frac{21}{64}\tau T(\langle \alpha^2 \rangle - \langle \alpha_O \alpha_V \rangle)$. [17]
Dans l'hypothèse (B), il faut remplacer $1+3\mathcal{R}_a$ par $1+\mathcal{T}_a$ et τ par T. Après élimination

16. Dans le cas $T = 30$ min : $\mathcal{T}_a^{\text{test}} \in [5,7; 6,1]$. Dans le cas $T = 1$h : $\mathcal{T}_a^{\text{test}} \in [1,0; 1,2]$.

17. Pour $T = 30$ min : $K_1 := \frac{3}{4}T\langle \alpha \rangle$, $K_2 := \frac{9}{16}T^2\langle \alpha_O \alpha_V \rangle$, $K_3 := \frac{5}{16}\tau T(\langle \alpha^2 \rangle - \langle \alpha_O \alpha_V \rangle)$ et $\langle P_a^2 \rangle_{\text{A}} = \frac{1}{12}K_1(7-3\mathcal{R}_a)\langle \overline{n_a} \rangle + K_3(1+3\mathcal{R}_a)\langle \overline{n_a} \rangle + K_2(1+2\mathcal{R}_a+\mathcal{S}_{aa})\langle (\overline{n_a})^2 \rangle$.
Pour $T = 1$h : $K_1 := \frac{1}{2}T\langle \alpha \rangle$, $K_2 := \frac{1}{4}T^2\langle \alpha_O \alpha_V \rangle$, $K_3 := \frac{1}{4}\tau T(\langle \alpha^2 \rangle - \langle \alpha_O \alpha_V \rangle)$ et $\langle P_a^2 \rangle_{\text{A}} = \frac{1}{2}K_1(1-\mathcal{R}_a)\langle \overline{n_a} \rangle + K_3(1+3\mathcal{R}_a)\langle \overline{n_a} \rangle + K_2(1+2\mathcal{R}_a+\mathcal{S}_{aa})\langle (\overline{n_a})^2 \rangle$.

des K_i on obtient pour les moments de n_O :

$$\begin{aligned}
\langle\overline{n_O}\rangle &= \left(\frac{1+\mathcal{R}_V}{1+\mathcal{R}_O}\right)\frac{\langle P_O\rangle}{\langle P_V\rangle}\langle\overline{n_V}\rangle \\
\langle(\overline{n_O})^2\rangle_{\mathrm{A}} &= \langle(\overline{n_V})^2\rangle\left(\frac{1+3\mathcal{R}_O}{1+\mathcal{R}_O}\right)\left(\frac{1+\mathcal{R}_V}{1+3\mathcal{R}_V}\right)\frac{\langle P_O\rangle}{\langle P_V\rangle}\left(\frac{1+2\mathcal{R}_V+\mathcal{S}_{VV}}{1+2\mathcal{R}_O+\mathcal{S}_{OO}}\right) \\
&\quad + \langle\overline{n_O}\,\overline{n_V}\rangle\left(\frac{1+\mathcal{R}_O+\mathcal{R}_V+\mathcal{S}_{OV}}{1+2\mathcal{R}_O+\mathcal{S}_{OO}}\right)\frac{1}{\langle P_O P_V\rangle}\left\{\langle P_O^2\rangle - \frac{1}{8}\left(\frac{5-\mathcal{R}_O}{1+\mathcal{R}_O}\right)\langle P_O\rangle\right. \\
&\quad \left. + \left(\frac{1+3\mathcal{R}_O}{1+\mathcal{R}_O}\right)\left(\frac{1+\mathcal{R}_V}{1+3\mathcal{R}_V}\right)\frac{\langle P_O\rangle}{\langle P_V\rangle}\left(\frac{1}{8}\left(\frac{5-\mathcal{R}_V}{1+\mathcal{R}_V}\right)\langle P_V\rangle - \langle P_V^2\rangle\right)\right\}
\end{aligned}$$

Dans l'hypothèse (B), il faut remplacer $1+3\mathcal{R}_a$ par $1+\mathcal{T}_a$. [18]

Remarque : pour $\mathcal{R}_O = \mathcal{R}_V$, $\mathcal{S}_{OO} = \mathcal{S}_{VV} = \mathcal{S}_{OV}$, et dans l'hypothèse (B) $\mathcal{T}_O = \mathcal{T}_V$, c'est-à-dire si les plasmides et le chromosome se répliquent et se répartissent de la même manière, on retrouve les formules du modèle simple : exprimer les moments de $\overline{n_O}$ en fonction de ceux de $\overline{n_V}$ permet de s'affranchir à la fois des fluctuations d'expression génétique et des fluctuations globales, seules restent ces différences « intrinsèques » entre plasmide et chromosome, dans l'esprit de [1].

B.8.2 Résultats

Le nombre de protéines EGFP ou mOrange n'est pas directement mesuré, mais seulement leur fluorescence, qui lui est proportionnelle. On ne peut donc pas prendre en compte la contribution de la répartition des protéines lors des divisions. Celle-ci est cependant vraisemblablement négligeable. Premièrement, le promoteur *pTacI* est très fort quand il est induit : plusieurs dizaines, voire plusieurs centaines de protéines sont produites lors de l'induction : $\langle P_a\rangle$ et certainement très inférieur à $\langle P_a^2\rangle$. Deuxièmement, cette contribution s'écrit (dans les accolades de l'expression précédente) :

$$\frac{\langle P_O\rangle}{8(1+\mathcal{R}_O)}\left(\left(\frac{1+3\mathcal{R}_O}{1+3\mathcal{R}_V}\right)(5-\mathcal{R}_V) - (5-\mathcal{R}_O)\right)$$

qui tend vers zéro pour $\mathcal{R}_O \sim \mathcal{R}_V$. Les résultats suivants seront donnés sans prendre en compte cette contribution.

Soit C_1 et C_2 les coefficients de $\langle(\overline{n_V})^2\rangle$ et $\langle\overline{n_O}\,\overline{n_V}\rangle$ dans l'expression de $\langle(\overline{n_O})^2\rangle$:

$$\langle(\overline{n_O})^2\rangle = C_1\langle(\overline{n_V})^2\rangle + C_2\langle\overline{n_O}\,\overline{n_V}\rangle$$

Le tableau B.1 présente les valeurs obtenues, après utilisation des données expérimentales, pour $\langle\overline{n_O}\rangle/\langle\overline{n_V}\rangle$, C_1 et C_2 dans les cas envisagés jusqu'ici : modèle simple, hyopthèse de

18. Pour des temps de division de 30 min et 1h seuls les termes provenant de la répartition des protéines lors des divisions diffèrent : il faut alors remplacer $(5-\mathcal{R}_a)/8$ par respectivement $(7-3\mathcal{R}_a)/12$ et $(1-\mathcal{R}_a)/2$.

« mémoire courte » (A), de « mémoire longue » (B), trois, deux, une divisions, avec ou sans fonctions tests, pour les quatre plasmides étudiés, à 30 °C. Pour l'estimation par les fonctions tests, la même fonction a été utilisée pour calculer $\mathcal{R}_a$ et $\mathcal{T}_a$, des valeurs (dix) appartenant à l'intervalle trouvé étant indépendamment attribuées à $\mathcal{S}_{ab}$. Dans le cas général, dix valeurs appartenant à leurs intervalles respectifs ont été indépendamment attribuées à $\mathcal{R}_O$, $\mathcal{R}_V$, $\mathcal{S}_{OO}$, $\mathcal{S}_{VV}$, $\mathcal{S}_{OV}$, $\mathcal{T}_O$ et $\mathcal{T}_V$ [19], ce qui conduit une nouvelle fois à des intervalles de taille très surestimée.

Je n'ai pas pris en compte les erreurs expérimentales.

Avec les fonctions tests, les intervalles trouvés sont petits et encadrent la valeur obtenue par le modèle simple ; les valeurs des coefficients C_1 et C_2 ne dépendent quasiment pas de l'hypothèse, (A) ou (B), sur le temps de corrélation d'expression, ni du nombre de divisions considérées. Dans le cas général les intervalles sont plus grands, avec typiquement un facteur 2 entre la borne inférieure et la borne supérieure. Ceux trouvés dans l'hypothèse (B), la moins réaliste, s'avèrent peu intéressants.

Comme indiqué à la fin du paragraphe B.2, on peut faire l'approximation d'une variance du nombre de copies de chromosome nulle ($\langle(\overline{n_V})^2\rangle \simeq \langle\overline{n_V}\rangle^2$) et que les nombres de copies de plasmide et de chromosome sont décorrélés ($\langle\overline{n_O}\ \overline{n_V}\rangle \simeq \langle\overline{n_O}\rangle\langle\overline{n_V}\rangle$). Le tableau B.2 présente les valeurs obtenues pour le bruit du nombre de copies de plasmide calculé dans ces hypothèses.

Avec les fonctions tests, on trouve des valeurs de bruit qui dépendent des plasmides, cohérente avec les valeurs attendues (environ 50% pour le pZC, un bruit plus petit pour le pOAR, qui possède un système de partition, que pour le pOU, un bruit relativement faible pour le pBR, dont le nombre moyen de copies est élevé). Dans le cas général, on ne peut rien dire sur les bruits. La température n'a d'effet visible sur le bruit dans aucun des cas envisagées.

Il existe cependant encore plusieurs façons d'améliorer ces résultats. Considérant que la réplication du chromosome est parfaitement contrôlée, on peut se restreindre, pour son nombre de copies, à utiliser des fonctions marche comme fonctions tests ; de plus, dans l'expression de $\mathcal{S}_{OV}$, on peut factoriser $\langle n_O n_V\rangle(t_1;t_2)$ en $\langle n_O\rangle(t_1)\langle n_V\rangle(t_2)$ et postuler la même forme pour chacune de ces deux fonctions que dans le calcul de $\mathcal{R}_a$ et $\mathcal{T}_a$. Les intervalles sur les bruits ont été obtenus en considérant tous les cas envisagés jusqu'ici ; on pourrait cependant légitimement ne considérer que l'hypothèse (A) d'une autocorrélation d'expression courte. Cela changerait peu les résultats obtenus avec des fonctions tests, mais réduirait de beaucoup la taille des intervalles trouvés dans le cas général. Enfin, si l'on veut comparer deux plasmides (par exemple calculer le rapport de leurs bruits), on peut faire un peu mieux en considérant que le chromosome se réplique de la même manière dans les deux cas. Si ces deux plasmides ont la même origine de réplication (ici, pOU et pOAR),

19. La borne inférieure de $1 + \mathcal{T}_a$ a été prise égale $0,01$ dans le cas $T = 1$h.

		pZC	pOU	pOAR	pBR
$\langle\overline{n_O}\rangle/\langle\overline{n_V}\rangle$	simple	0,47	6,2	4,6	76
	test	[0,45 ; 0,50]	[5,9 ; 6,6]	[4,3 ; 4,9]	[70 ; 79]
	général	[0,38 ; 0,60]	[4,9 ; 7,9]	[3,6 ; 5,8]	[59 ; 93]
C_1 simple		0,47	6,2	4,6	76
C_1 (A) test		[0,44 ; 0,51]	[5,8 ; 6,7]	[4,3 ; 4,9]	[69 ; 79]
C_1 (B) test	3 div.	[0,42 ; 0,54]	[5,5 ; 7,0]	[4,0 ; 5,2]	[65 ; 84]
	2 div.	[0,42 ; 0,54]	[5,5 ; 7,1]	[4,0 ; 5,2]	[65 ; 84]
	1 div.	[0,44 ; 0,51]	[5,8 ; 6,7]	[4,2 ; 4,9]	[68 ; 80]
C_1 (A) général		[0,34 ; 0,67]	[4,4 ; 8,8]	[3,3 ; 6,4]	[53 ; 104]
C_1 (B) général	3 div.	[0,03 ; 6,5]	[0,46 ; 85]	[0,34 ; 62]	[5,4 ; 1000]
	2 div.	[0,02 ; 12]	[0,25 ; 155]	[0,18 ; 114]	[3,0 ; 1800]
	1 div.	[0,0 ; 480]	[0,01 ; 6250]	[0,0 ; 4600]	[0,1 ; 74000]
C_2 simple		-0,39	6,2	4,0	79
C_2 (A) test		[-0,45 ; -0,34]	[5,7 ; 6,8]	[3,6 ; 4,4]	[72 ; 84]
C_2 (B) test	3 div.	[-0,44 ; -0,36]	[5,8 ; 6,8]	[3,7 ; 4,3]	[72 ; 84]
	2 div.	[-0,44 ; -0,35]	[5,8 ; 6,8]	[3,7 ; 4,3]	[72 ; 84]
	1 div.	[-0,45 ; -0,34]	[5,7 ; 6,8]	[3,6 ; 4,4]	[72 ; 84]
C_2 (A) général		[-0,72 ; -0,16]	[4,0 ; 10]	[2,5 ; 6,6]	[52 ; 122]
C_2 (B) général	3 div.	[-11 ; 0,6]	[-0,72 ; 11]	[-4,3 ; 7,5]	[50 ; 123]
	2 div.	[-20 ; 0,7]	[-11 ; 11]	[-14 ; 7,7]	[47 ; 124]
	1 div.	[-850 ; 0,9]	[-880 ; 11]	[-890 ; 7,9]	[-780 ; 124]

Tab. B.1 – Valeurs obtenues pour les coefficients des moments de $\overline{n_O}$, dans différents cas, pour des expériences à 30 ° C. Dans l'hypothèse (B), le cas « 1 division » est le plus réaliste.

Temp.	Cas	pZC	pOU	pOAR	pBR
30 ° C	simple	0,55	0,39	0,29	0,25
	test	[0,37 ; 0,66]	[0,31 ; 0,48]	[0,14 ; 0,39]	[0,02 ; 0,35]
	général	[0 ; 1,2]	[0 ; 0,78]	[0 ; 0,71]	[0 ; 0,66]
32 ° C	simple	0,47	0,34	0,28	0,24
	test	[0,33 ; 0,58]	[0,24 ; 0,44]	[0,14 ; 0,39]	[0 ; 0,34]
	général	[0 ; 1,1]	[0 ; 0,74]	[0 ; 0,71]	[0 ; 0,66]
35 ° C	simple	0,46	0,36	0,28	0,21
	test	[0,34 ; 0,59]	[0,26 ; 0,45]	[0,14 ; 0,39]	[0 ; 0,33]
	général	[0 ; 1,1]	[0 ; 0,75]	[0 ; 0,70]	[0 ; 0,65]

Tab. B.2 – Valeurs du bruit du nombre moyen sur un cycle de copies de plasmide calculées dans l'hypothèse où la variance du nombre de copies de chromosome est nulle et les nombres de copies de plasmide et de chromosome sont décorrélés. Dans le cas général, l'hypothèse « 1 division » n'a pas été prise en compte.

on peut de plus supposer que $\mathcal{R}_O$ et $\mathcal{T}_O$ prennent les mêmes valeurs pour chacun.

B.9 Conclusion

Moyennant des hypothèses sur l'expression génétique et la réplication du chromosme, nous avons pu extraire des données de fluorescence les bruits de nombres de copies de plasmides, ou plus exactement des nombres moyens sur un cycle. Une incertitude provenant de notre ignorance sur la façon dont les plasmides et le chromosome se répliquent et se répartissent à la division demeure. L'étude dans le cas général apporte peu d'enseignement sur les fluctuations de nombre de copies, mais elle surestime fortement cette incertitude. Un travail d'analyse plus poussé améliorerait certainement beaucoup les résultats.

Considérer un jeu de fonctions tests de profils variés donne des estimations bien meilleures. Si les résultats présentés peuvent encore être légèrement améliorés, on peut déjà distinguer les comportements de plasmides différents et notamment observer qu'un système de partition diminue bien le bruit de nombre de copies.

Il est frappant de constater que le modèle simple introduit au début de cette étude conduit à des valeurs comprises dans les intervalles trouvés en considérant toutes les sources de fluctuation : le fait d'utiliser le chromosome comme référence et de mesurer dans une même bactérie les fluorescences verte et orange permet de s'affranchir des perturbations affectant de la même manière les plasmides et le chromosome (fluctuations d'expression, durée du cycle ou nombre de divisions, dans une large mesure la répartition des protéines fluorescentes à la division). On peut ainsi penser que l'hypothèse que le temps d'induction est un multiple du temps de division, juste pour les températures de 30 et 37 ° C mais

erronée pour des températures intermédiaires, n'induit pas une erreur importante sur les résultats.

Connaître le bruit du nombre de plasmides constituera un apport important aux nombreux travaux sur le bruit d'expression génétique où ils sont utilisés et permettra de mieux comprendre les réseaux de régulation qui le contrôlent. Enfin, il sera possible d'estimer si, en variant la pression de sélection (quantité d'antibiotique dont les plasmides portent la résistance par exemple), la variabilité du nombre de copies de plasmide peut constituer un facteur d'adaptabilité.

Bibliographie

[1] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584) :1183–1186, Aug 2002.

[2] Johan Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973) :415–418, Jan 2004.

[3] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance : stochastic gene expression and its consequences. *Cell*, 135(2) :216–226, Oct 2008.

[4] Christopher V Rao, Denise M Wolf, and Adam P Arkin. Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912) :231–237, Nov 2002.

[5] Ertugrul M Ozbudak, Mukund Thattai, Iren Kurtser, Alan D Grossman, and Alexander van Oudenaarden. Regulation of noise in the expression of a single gene. *Nat Genet*, 31(1) :69–73, May 2002.

[6] Nitzan Rosenfeld, Jonathan W Young, Uri Alon, Peter S Swain, and Michael B Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717) :1962–1965, Mar 2005.

[7] N. Barkai and S. Leibler. Robustness in simple biochemical networks. *Nature*, 387(6636) :913–917, Jun 1997.

[8] U. Alon, M. G. Surette, N. Barkai, and S. Leibler. Robustness in bacterial chemotaxis. *Nature*, 397(6715) :168–171, Jan 1999oz.

[9] Juan M Pedraza and Alexander van Oudenaarden. Noise propagation in gene networks. *Science*, 307(5717) :1965–1969, Mar 2005.

[10] Mary J Dunlop, Robert Sidney Cox, Joseph H Levine, Richard M Murray, and Michael B Elowitz. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nat Genet*, 40(12) :1493–1498, Dec 2008.

[11] Amos B Oppenheim, Oren Kobiler, Joel Stavans, Donald L Court, and Sankar Adhya. Switches in bacteriophage lambda development. *Annu Rev Genet*, 39 :409–429, 2005.

[12] Mark Ptashne. *A Genetic Switch, Phage λ and Higher Organisms.* Cell Press & Blackwell Publishing, 2nde edition, 1992.

[13] Larry Snyder and Wendy Champness. *Molecular Genetics of Bacteria.* ASM Press, 2nde edition, 2003.

[14] Sherwood R Casjens and Roger W Hendrix. Bacteriophage lambda and relatives. *Encyclopedia of Life Sciences (www.els.net)*, 2001.

[15] Oren Kobiler, Assaf Rokney, Nir Friedman, Donald L Court, Joel Stavans, and Amos B Oppenheim. Quantitative kinetic analysis of the bacteriophage lambda genetic network. *Proc Natl Acad Sci U S A*, 102(12) :4470–4475, Mar 2005.

[16] François St-Pierre and Drew Endy. Determination of cell fate selection during phage lambda infection. *Proc Natl Acad Sci U S A*, 105(52) :20705–20710, Dec 2008.

[17] P. J. Darling, J. M. Holt, and G. K. Ackers. Coupled energetics of lambda cro repressor self-assembly and site-specific dna operator binding ii : cooperative interactions of cro dimers. *J Mol Biol*, 302(3) :625–638, Sep 2000.

[18] Ian B Dodd, Keith E Shearwin, and J. Barry Egan. Revisited gene regulation in bacteriophage lambda. *Curr Opin Genet Dev*, 15(2) :145–152, Apr 2005.

[19] Sébastien Lemire. *Les prophages de Salmonella Typhimurium : Régulation lysogénique et contribution à la pathogénicité*. PhD thesis, Université Paris Sud-Orsay (Paris 11), Orsay, 2006.

[20] Shota Atsumi and John W Little. Role of the lytic repressor in prophage induction of phage lambda as analyzed by a module-replacement approach. *Proc Natl Acad Sci U S A*, 103(12) :4558–4563, Mar 2006.

[21] Sivan Pearl, Chana Gabay, Roy Kishony, Amos Oppenheim, and Nathalie Q Balaban. Nongenetic individuality in the host-phage interaction. *PLoS Biol*, 6(5) :e120, May 2008.

[22] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected escherichia coli cells. *Genetics*, 149(4) :1633–1648, Aug 1998.

[23] José M G Vilar and Stanislas Leibler. Dna looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5) :981–989, Aug 2003.

[24] Marco J Morelli, Pieter Rein Ten Wolde, and Rosalind J Allen. Dna looping provides stability and robustness to the bacteriophage lambda switch. *Proc Natl Acad Sci U S A*, 106(20) :8101–8106, May 2009.

[25] C. S. Shean and M. E. Gottesman. Translation of the prophage lambda cl transcript. *Cell*, 70(3) :513–522, Aug 1992.

[26] Isabella Moll, Go Hirokawa, Michael C Kiel, Akira Kaji, and Udo BlÃ¤si. Translation initiation with 70s ribosomes : an alternative pathway for leaderless mrnas. *Nucleic Acids Res*, 32(11) :3354–3363, 2004.

[27] Sine L Svenningsen, Nina Costantino, Donald L Court, and Sankar Adhya. On the role of cro in lambda prophage induction. *Proc Natl Acad Sci U S A*, 102(12) :4465–4469, Mar 2005.

[28] Assaf Rokney, Oren Kobiler, Amnon Amir, Donald L Court, Joel Stavans, Sankar Adhya, and Amos B Oppenheim. Host responses influence on the induction of lambda prophage. *Mol Microbiol*, 68(1) :29–36, Apr 2008.

[29] G. Plunkett and H. Echols. Retroregulation of the bacteriophage lambda int gene : limited secondary degradation of the rnase iii-processed transcript. *J Bacteriol*, 171(1) :588–592, Jan 1989.

[30] Amnon Amir, Oren Kobiler, Assaf Rokney, Amos B Oppenheim, and Joel Stavans. Noise in timing and precision of gene activities in a genetic cascade. *Mol Syst Biol*, 3 :71, 2007.

[31] Erik Aurell, Stanley Brown, Johan Johanson, and Kim Sneppen. Stability puzzles in phage lambda. *Phys Rev E Stat Nonlin Soft Matter Phys*, 65(5 Pt 1) :051914, May 2002.

[32] X-M. Zhu, L. Yin, L. Hood, and P. Ao. Calculating biological behaviors of epigenetic states in the phage lambda life cycle. *Funct Integr Genomics*, 4(3) :188–195, Jul 2004.

[33] Tianhai Tian and Kevin Burrage. Bistability and switching in the lysis/lysogeny genetic regulatory network of bacteriophage lambda. *J Theor Biol*, 227(2) :229–237, Mar 2004.

[34] Daniel Schultz, Aleksandra M Walczak, José N Onuchic, and Peter G Wolynes. Extinction and resurrection in gene networks. *Proc Natl Acad Sci U S A*, 105(49) :19165–19170, Dec 2008.

[35] Kristoffer Baek, Sine Svenningsen, Harvey Eisen, Kim Sneppen, and Stanley Brown. Single-cell analysis of lambda immunity regulation. *J Mol Biol*, 334(3) :363–372, Nov 2003.

[36] L. Meadow Anderson and Haw Yang. Dna looping can enhance lysogenic ci transcription in phage lambda. *Proc Natl Acad Sci U S A*, 105(15) :5827–5832, Apr 2008.

[37] Hiizu Nakanishi, Namiko Mitarai, and Kim Sneppen. Dynamical analysis on gene activity in the presence of repressors and an interfering promoter. *Biophys J*, 95(9) :4228–4240, Nov 2008.

[38] M. Lieb. Studies of heat-inducible lambda bacteriophage. i. order of genetic sites and properties of mutant prophages. *J Mol Biol*, 16(1) :149–163, Mar 1966.

[39] N. K. Jana, S. Roy, B. Bhattacharyya, and N. C. Mandal. Amino acid changes in the repressor of bacteriophage lambda due to temperature-sensitive mutations in its ci gene and the structure of a highly temperature-sensitive mutant repressor. *Protein Eng*, 12(3) :225–233, Mar 1999.

[40] S. Naono and F. Gros. On the mechanism of transcription of the lambda genome during induction of lysogenic bacteria. *J Mol Biol*, 25(3) :517–536, May 1967.

[41] S. M. Uptain and M. J. Chamberlin. Escherichia coli rna polymerase terminates transcription efficiently at rho-independent terminators on single-stranded dna templates. *Proc Natl Acad Sci U S A*, 94(25) :13548–13553, Dec 1997.

[42] Nathan C Shaner, Michael Z Lin, Michael R McKeown, Paul A Steinbach, Kristin L Hazelwood, Michael W Davidson, and Roger Y Tsien. Improving the photostability of bright monomeric orange and red fluorescent proteins. *Nat Methods*, 5(6) :545–551, Jun 2008.

[43] Nathan C Shaner, Paul A Steinbach, and Roger Y Tsien. A guide to choosing fluorescent proteins. *Nat Methods*, 2(12) :905–909, Dec 2005.

[44] Judith A Megerle, Georg Fritz, Ulrich Gerland, Kirsten Jung, and Joachim O Rädler. Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophys J*, 95(4) :2103–2115, Aug 2008.

[45] R. Y. Tsien. The green fluorescent protein. *Annu Rev Biochem*, 67 :509–544, 1998.

[46] Jeffrey R Chabot, Juan M Pedraza, Prashant Luitel, and Alexander van Oudenaarden. Stochastic gene expression out-of-steady-state in the cyanobacterial circadian clock. *Nature*, 450(7173) :1249–1252, Dec 2007.

[47] J. Shi and D. P. Biek. A versatile low-copy-number cloning vector derived from plasmid f. *Gene*, 164(1) :55–58, Oct 1995.

[48] Daniel Sinnecker, Philipp Voigt, Nicole Hellwig, and Michael Schaefer. Reversible photobleaching of enhanced green fluorescent proteins. *Biochemistry*, 44(18) :7085–7094, May 2005.

[49] Gilles Charvin, Frederick R Cross, and Eric D Siggia. A microfluidic device for temporally controlled gene expression and long-term fluorescent imaging in unperturbed dividing yeast cells. *PLoS One*, 3(1) :e1468, 2008.

[50] P. Kourilsky. Lysogenization by bacteriophage lambda. i. multiple infection and the lysogenic response. *Mol Gen Genet*, 122(2) :183–195, Apr 1973.

[51] Shota Atsumi and John W Little. A synthetic phage lambda regulatory circuit. *Proc Natl Acad Sci U S A*, 103(50) :19045–19050, Dec 2006.

[52] Paul François. *Réseaux Génétiques : Conception, Modélisation et Dynamique*. PhD thesis, École Normale Supérieure - Université Pierre et Marie Curie (Paris 6), Paris, 2005.

[53] Paul François and Vincent Hakim. Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci U S A*, 101(2) :580–585, Jan 2004.

[54] J. N. Weiss. The hill equation revisited : uses and misuses. *FASEB J*, 11(11) :835–841, Sep 1997.

[55] Thomas Kuhlman, Zhongge Zhang, Milton H Saier, and Terence Hwa. Combinatorial transcriptional control of the lactose operon of escherichia coli. *Proc Natl Acad Sci U S A*, Mar 2007.

[56] J. Reinitz and J. R. Vaisnys. Theoretical and experimental analysis of the phage lambda genetic switch implies missing levels of co-operativity. *J Theor Biol*, 145(3) :295–318, Aug 1990.

[57] Paul François, Vincent Hakim, and Eric D. Siggia. Deriving structure from evolution : metazoan segmentation. *Mol Syst Biol*, 3 :154, 2007.

[58] Paul François and Vincent Hakim. Core genetic module : the mixed feedback loop. *Phys Rev E Stat Nonlin Soft Matter Phys*, 72(3 Pt 1) :031908, Sep 2005.

[59] Thomas Blumenthal. Operons in eukaryotes. *Brief Funct Genomic Proteomic*, 3(3) :199–211, Nov 2004.

[60] Joseph Sambrook and David W Russel. *Molecular Cloning.* Cold Spring Harbor Laboratory Press, 3ème edition, 2001.

[61] Jérôme Wong Ng. *Variation du Nombre de Copies de Plasmides au Sein de Populations Monoclonales de Bactéries.* PhD thesis, École Normale Supérieure - Université Pierre et Marie Curie (Paris 6), Paris, 2008.

[62] Kurt Nordström and Santanu Dasgupta. Copy-number control of the escherichia coli chromosome : a plasmidologist's view. *EMBO Rep*, 7(5) :484–489, May 2006.

[63] Eric J Stewart, Richard Madden, Gregory Paul, and François Taddei. Aging and death in an organism that reproduces by morphologically symmetric division. *PLoS Biol*, 3(2) :e45, Feb 2005.

[64] Frederik C. Neidhart, editor. *Escherichia coli and Salmonella, Cellular and Molecular Biology.* ASM Press, 2ème edition, 1996.

Remerciements

Je remercie chaleureusement Alice Aubin, Juliette Ben Arous, Didier Chatenay, Cyril Cichowlas, Elie Desmond, Amor Ghozzi, Sabine Ghozzi, Jérôme Robert, Samuel Rondière et Jérôme Wong Ng pour leur soutien, leurs encouragements, leur indulgence et leur patience.

Je remercie Richard d'Ari de m'avoir initié à Lambda et fait partager sa passion, pour ses conseils précieux et ses encouragements.

Merci encore à Didier Chatenay, Jérôme Robert et Jérôme Wong Ng pour leur aide constante et leurs conseils.

Je remercie Paul François, Vincent Hakim et Hervé Rouault de m'avoir fourni le code de *Genherite*, et plus particulièrement Hervé pour son aide en programmation et sur la modélisation des réseaux de régulation.

J'ai beaucoup appris de Marc Dreyfus, Max Gottesman, Sébastien Lemire, Mathias Springer et Eric Stewart ; je les remercie d'avoir pris le temps de discuter de ce travail et de leurs suggestions.

Je remercie Simona Cocco de m'avoir aidé à obtenir une allocation de thèse.

Je remercie les membres de mon jury d'avoir accepté d'en faire partie et tout particulièrement Bahram Houchmandzadeh et Denis Thieffry d'avoir bien voulu être rapporteurs.

Merci enfin à Marie Gefflot, Annie Ribaudeau et Nora Sadaoui pour leur bonne volonté, leur aide et leur amabilité.

Inference of plasmid copy number mean and noise from single cell gene expression data

Stéphane Ghozzi* and Jérôme Wong Ng†
Laboratoire de Physique Statistique, École Normale Supérieure, UPMC Univ Paris 06, Université Paris Diderot, CNRS, 24 rue Lhomond, 75005 Paris, France.

Didier Chatenay and Jérôme Robert
Laboratoire Jean Perrin, FRE 3231 CNRS-UPMC, 24 rue Lhomond, 75005 Paris, France.
(Dated: August 26, 2010)

Plasmids are extra-chromosomal DNA molecules which code for their own replication. We previously reported a setup using genes coding for fluorescent proteins of two colors that allowed us, using a simple model, to extract the plasmid copy number noise in a monoclonal population of bacteria [J. Wong Ng et al., Phys. Rev. E, 81, 011909 (2010)]. Here we present a detailed calculation relating this noise to the measured levels of fluorescence, taking into account all sources of fluorescence fluctuations: the fluctuation of gene expression as in the simple model, but also the growth and division of bacteria, the non-uniform distribution of their ages, the random partition of proteins at divisions and the replication and partition of plasmids and chromosome. We show how using the chromosome as a reference helps extracting the plasmid copy number noise in a self-consistent manner.

PACS numbers: 87.18.Tt, 87.16.-b
Keywords: Plasmid copy number; stochastic gene expression; phenotypic variability

I. INTRODUCTION

Plasmids are extra-chromosomal DNA molecules which code for their own replication [1]. They are highly common in natural bacterial strains and are widely used in studies of gene expression; they have been seen as a model for genomic replication and partition [2] and studied as genetic control systems [3].

In a previous article, we described an experiment in which we collected fluorescence signals from modified plasmids in a monoclonal population of bacteria and used the chromosome as a reference to get insight in the plasmid copy number (PCN) distributions [4]. Here we show how the measured levels of fluorescence relate to the PCN mean and noise.

We briefly recall the setup of the experiment in the remainder of this Introduction. In Section II we derive the expression for PCN mean and noise in a simple case, where only fluctuations of gene expression are considered. The realistic case, taking into account all sources of fluctuations of the actual experiment, is presented in Section III. Section IV presents the values obtained for PCN mean and noise when one uses the experimentally measured quantities. These results and the principle of this work are then discussed. Appendixes present some computations in greater details.

The gene *egfp* [5], coding for the the green fluorescent protein EGFP, was fused to the inducible, strong promoter *PtacI* [6] and then inserted in the chromosome of an *E. coli* strain. The bacteria were then transformed with either one of the four plasmids studied here, which contained the fusion *PtacI-mOrange* [7]: we thus obtained strains expressing EGFP and the orange fluorescent protein mOrange at the same time, under the same transcriptional control. After one hour induction with IPTG, all protein expression was blocked. Cells were incubated overnight so that all fluorescent proteins acquire their mature form. For each of the four strains, green and orange fluorescence intensities of individual cells were then measured. In each experiment at least 10,000 cells were observed, and at least three experiments were done in each condition.

In general, disentangling the various contributions to the final distribution of fluorescence would be a difficult problem. However, making some assumptions on the gene expression processes, we will be able to express the first and second moments of the number of fluorescent proteins as functions of those of copy numbers and to inverse these relations to find how to relate the experimental measurements to the distribution of PCN. The next section presents this strategy in a simple case.

II. SIMPLE MODEL

We suppose here that during the induction, bacteria do not grow, the plasmids and chromosomes do not replicate, the protein production does not depend on time [8] and the age distribution of bacteria is uniform.

We note P_a^i the contribution of the copy i of the gene a ($a = O$ or G for the genes *mOrange* or *egfp*) to the total number of proteins P_a at the end of induction in one cell and n_a the number of copies of the gene a in that cell

* now at Institut für Theoretische Physik, Universität zu Köln, Zülpicherstrasse 77, 50937 Cologne, Germany; ghozzi@thp.uni-koeln.de
† now at Physics of Biological Systems, CNRS URA 2171, Institut Pasteur, 25-28, rue du Dr Roux, 75015 Paris, France.

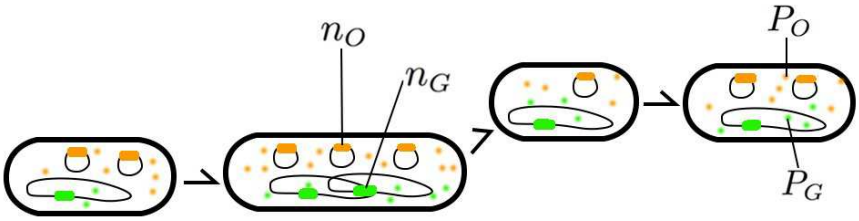


FIG. 1. (Color online) Cartoon of the lineage of a bacterium during protein production induction, here depicted with one division (only one of the two final cells is shown). Fluorescence intensities of single cells are measured at the end of induction. The orange (resp. green) intensities are proportional to the number of orange proteins P_O (resp. green proteins P_G) in the observed cell. These proteins were produced during all the induction by a varying number of *mOrange* or *egfp* copies (n_O and n_G) and randomly distributed among daughter cells at each division.

(see Fig. 1). One can write:

$$P_a = \sum_{i=1}^{n_a} P_a^i.$$

The average (over the population) of P_a can thus be written:

$$\langle P_a \rangle = \sum_{n_a} \sum_{i=1}^{n_a} \sum_{P_a^i} p(n_a, P_a^i)\, P_a^i,$$

where $p(n_a, P_a^i)$ is the joint probability of n_a and P_a^i. We can suppose that the distribution of the number of proteins produced by each copy does not depend on the particular copy considered nor on the number of copies (we measured the same distributions of green fluorescence, i.e. of expression from the chromosome, for strains bearing both high and low copy number plasmids [9]). Thus:

$$\begin{aligned}\langle P_a \rangle &= \sum_{n_a} p(n_a)\, n_a \sum_{P_a^1} p(P_a^1)\, P_a^1 \\ &= \langle n_a \rangle \langle P_a^1 \rangle.\end{aligned}$$

Moreover we can suppose that on average the number of proteins produced by a copy of a gene does not depend on the gene (both genes are under the same promoter). Hence, as expected:

$$\frac{\langle n_O \rangle}{\langle n_G \rangle} = \frac{\langle P_O \rangle}{\langle P_G \rangle}. \tag{1}$$

The moments of order 2 can similarly be written:

$$\langle P_a P_b \rangle = \sum_{n_a, n_b} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} \sum_{P_a^i, P_b^j} p(n_a, n_b, P_a^i, P_b^j)\, P_a^i P_b^j.$$

where P_a and P_b are evaluated in the same cell.

In the case of different genes, we can suppose that the correlation does not depend on the particular copies considered, nor on their numbers. Thus:

$$\begin{aligned}\langle P_O P_G \rangle &= \sum_{n_O, n_G} p(n_O, n_G)\, n_O n_G \sum_{P_O^1, P_G^1} p(P_O^1, P_G^1)\, P_O^1 P_G^1 \\ &= \langle n_O n_G \rangle \langle P_O^1 P_G^1 \rangle.\end{aligned}$$

In the case of the same gene, we can suppose that two different copies correlate like two copies of different genes ($\langle P_a^i P_a^j \rangle = \langle P_O^1 P_G^1 \rangle,\ \forall i \neq j$) and that the autocorrelation of one copy does not depend on the particular copy or gene considered ($\langle (P_a^i)^2 \rangle = \langle (P^1)^2 \rangle,\ \forall a, i$). Then:

$$\langle P_a^2 \rangle = \langle n_a \rangle \langle (P^1)^2 \rangle + \langle n_a (n_a - 1) \rangle \langle P_O^1 P_G^1 \rangle.$$

Combining those two last expressions with equation 1, we obtain:

$$\langle n_O^2 \rangle = \frac{\langle P_O \rangle}{\langle P_G \rangle} \langle n_G^2 \rangle + \frac{1}{\langle P_O P_G \rangle} \left(\langle P_O^2 \rangle - \frac{\langle P_O \rangle}{\langle P_G \rangle} \langle P_G^2 \rangle \right) \langle n_O n_G \rangle.$$

Since the replication of the chromosome is well controlled [2, 10] we can suppose that the variance of the chromosome copy number vanishes ($\langle n_G^2 \rangle \simeq \langle n_G \rangle^2$) and that the plasmid and chromosome copy numbers are uncorrelated ($\langle n_O n_G \rangle \simeq \langle n_O \rangle \langle n_G \rangle$). Let η be the PCN noise, which is defined by: $\eta^2 = (\langle n_O^2 \rangle - \langle n_O \rangle^2) / \langle n_O \rangle^2$. Then:

$$\eta^2 = \frac{\langle P_G \rangle}{\langle P_O \rangle} + \frac{1}{\langle P_O P_G \rangle} \left(\frac{\langle P_G \rangle}{\langle P_O \rangle} \langle P_O^2 \rangle - \langle P_G^2 \rangle \right) - 1, \tag{2}$$

which, it turns out, does not depend on the chromosome copy number or any other external inputs, but solely on quantities directly measured in this experiment.

III. COMPLETE MODEL

We want now to also take into account sources of fluorescence fluctuation other than gene expression. We assume that *all cells have exactly the same division time* T. Two studies report a small variability of division times, with a standard deviation of the growth time constant of ∼10% of the average [11, 12]. We performed the induction on bulk, agitated cultures, which we expect to experience less growth variability than cells grown on solid substrates.

We note t_0 the age of a cell at the beginning of induction; under this hypothesis, the distribution of ages t_0 is

exponential [13]: $p(t_0) = (2\ln 2/T).2^{-t_0/T}$. We will also assume that the induction time (one hour) is a multiple of the division time. This is true at 30 and 37 °C, where we measured cell cycles of 1 h and 30 min respectively, but not for intermediate temperatures (this is discussed in Section IV). We will present calculations assuming that cells divide twice during the induction, i.e. a cell cycle of 30 min; more or less divisions only change the numerical pre-factors [14].

At each cell division, fluorescent proteins are randomly inherited by one of the two daughter cells, thus adding to the fluorescence fluctuations. As discussed in the Appendixes, this contribution turns out to be small. We will assume now that exactly half of the fluorescent proteins are inherited by each daughter cell.

Following one lineage during the induction, we can now express the number of fluorescent proteins at the end of induction in a given cell:

$$P_a = \left(\frac{1}{4}\int_{t_0}^{T} + \frac{1}{2}\int_{T}^{2T} + \int_{2T}^{2T+t_0}\right)\sum_{i=1}^{n_a(t)} \alpha_a(i,t)\,dt,$$

where we took the age of the cell at the beginning of induction t_0 as the initial time and introduced $\alpha_a(i,t)$, the rate of protein production at time t from the copy i of the gene a [15].

A. Fluorescence averages

To compute the average of P_a we introduce the joint probability $p[t_0, n_a, \alpha_a]$, which is now a functional and the integral is performed over all possible n_a and α_a functions:

$$\langle P_a\rangle = \int dt_0 \mathcal{D}[n_a]\mathcal{D}[\alpha_a]\; p[t_0, n_a, \alpha_a]\; P_a[t_0, n_a, \alpha_a].$$

We suppose that *the age at the beginning of induction t_0, the copy numbers of plasmid or chromosome n_a, the rates of protein production α_a are independent:* the above probability factorizes. This means that there is no growth or expression burden associated with the presence of the plasmids. The notion is still debated, and whereas we could extract a small inhibition of growth for the strain bearing the mini-R1-*par*$^-$ plasmid, we did not see any systematic deviation: we measured comparable growth rates for all strains at a given temperature; moreover, all strains exhibited the same average green fluorescence (thus, the same average protein production) [9].

We suppose also that during the induction, *the average protein production rate does not depend on time, on the particular copy considered or on the gene,* egfp *or* mOrange; similarly, *the correlations of two different copies do not depend on time, on the copies or on the genes.* We assume that the protein production rates immediately reach a stationary state (but this has not to hold for the proteins concentrations): Elf *et al.* have shown that, at 1 mM IPTG, the fraction of LacI bound to the Lac promoter reaches its steady state value (zero) in less than 10 s [16]. Whereas the promoter we used, *PtacI*, is slightly different, there is no reason for the dynamics to be slower [17].

We assume that the copies of a given gene are undistinguishable.

We assume that the dynamics of expression of *egfp* and *mOrange* are essentially fixed by the promoter; since it is the same for both genes, any copy of any of the two will follow the same statistics. Any systematic difference in the translation rate (mRNA lifetime, codon usage) is incorporated in the fluorescence per molecule factor. Since the cells are incubated overnight, with chloramphenicol blocking protein production, we expect all fluorescent proteins to have acquired their mature form. Lastly, both genes showed the same distribution of fluorescence when inserted in the chromosome [4].

We assume no dependance of the average expression on the time in the cell cycle.

Similarly, we assume that the gene copy numbers have reached a steady state and that there are no active loss of plasmid during the cell cycle or systematic bias in the way plasmids are inherited by daughter cells upon division: on average the plasmid and chromosome copy numbers are periodic of period T and $\langle n_a(T)\rangle = 2\langle n_a(0)\rangle$.

We then find:

$$\langle P_a\rangle = \frac{3}{4}\,T\langle\alpha\rangle\left(\langle\overline{n_a}\rangle + \frac{1}{T}\int_0^T dt_0\, p(t_0)\int_0^{t_0} dt\,\langle n_a(t)\rangle\right),$$

where $\overline{\bullet}$ is the average over one cycle, which commutes with the average over the population.

In general we cannot inverse this relation so as to express the average copy number as a function of the average protein number and we do not know the plasmid replication systems well enough to evaluate the second term in the parentheses. It is nevertheless possible to bound its ratio to the mean copy number. We thus define $\mathcal{R}_a = ((1/T)\int_0^T dt_0\, p(t_0)\int_0^{t_0} dt\,\langle n_a\rangle(t))/\langle\overline{n_a}\rangle$, and use it to express the mean PCN per chromosome:

$$\frac{\langle\overline{n_O}\rangle}{\langle\overline{n_G}\rangle} = \left(\frac{1+\mathcal{R}_G}{1+\mathcal{R}_O}\right)\frac{\langle P_O\rangle}{\langle P_G\rangle}. \tag{3}$$

We show in the Appendixes that $\mathcal{R}_a \in [0.15, 0.45]$. We also computed it after postulating various shapes for $\langle n_a\rangle$ as a function of time and propose that this interval can be reduced to $[0.36, 0.44]$ (see the Appendixes). The results for the four plasmids we studied, at various temperatures, are presented in Section IV.

B. Fluorescence cross-correlations

We will follow the same strategy for the correlations, namely bound terms related to plasmid or chromosome replication and partition. As in Section II, we suppose

that on average the cross-correlations of the rates of production of proteins and of the copy numbers of two different genes do not depend on the particular copies considered nor on time for the first (we discuss different forms of the rates auto-correlation below and show that they do not affect our results much), and that the chromosome replication and partition are perfectly controlled.

We introduce

$$\mathcal{S}_{ab} := \frac{1}{\langle \overline{n_a}\, \overline{n_b} \rangle} \frac{1}{T^2} \int_0^T dt_0\, p(t_0) \int_0^{t_0} dt \int_0^{t_0} dt' \langle n_a(t) n_b(t') \rangle,$$

and can now write:

$$\langle P_O P_G \rangle = \frac{9}{16} T^2 \langle \alpha_O \alpha_G \rangle (1 + \mathcal{R}_O + \mathcal{R}_G + \mathcal{S}_{OG}) \langle \overline{n_O} \rangle \langle \overline{n_G} \rangle, \tag{4}$$

where P_O and P_G are evaluated in the same cell. We show in the Appendixes that $\mathcal{S}_{OG} \in [0, 0.45]$, and argue that this interval can be reduced to $[0.20, 0.28]$.

C. Fluorescence auto-correlations

We consider now the moment of order 2 for the same gene, i.e. $\langle P_a^2 \rangle$, with $a = O$ or G. We approximate *the plasmid copy number auto-correlation function by a constant:*

$$\langle n_O(t) n_O(t') \rangle - \langle n_O(t) \rangle \langle n_O(t') \rangle = C_{n_O}\ , \forall t, t'.$$

This implies that $\langle n_O(t) n_O(t') \rangle$ is periodic in each of its arguments and allows us to transform the integrals over the induction time to integrals over one cell cycle. In the absence of a consensus model for plasmid replication or independent measurements, we cannot gauge *a priori* the error we thus make. Note however that C_{n_O} will not appear in the result.

During the induction, *the auto-correlation of the expression of a given copy does not depend on the gene considered, on the particular copy or on the time, but solely on the difference between two times:*

$$\langle \alpha_a(i,t) \alpha_a(i',t') \rangle - \langle \alpha \rangle^2 = C_\alpha(|t - t'|),$$

where i' is the *ancestor* of i. This follows from the same arguments as given above (e.g. the dependency on $|t - t'|$ follows from the assumption that the rate of protein production reached its steady state). We assume that the rates of expression of two different copies of a given gene correlate like those of two copies of two different genes.

Our guess is that the results will not be affected by the particular form this auto-correlation function will take; to test it we will make two extreme hypotheses: (A) of very short "memory", (B) of infinite (over the whole induction time) "memory".

In the hypothesis (A), we suppose that after a very short time τ the expression of a copy of a gene correlates with itself the same way it correlates with other copies. This makes sense if τ is small compared to the replication time; and indeed, we expect a particular copy auto-correlation to stem from multiple translations of a given mRNA, which has a typical life time of the order of the minute in bacteria, or from transcriptional bursts, which were shown to happen over short time scales [18]. In contrast, genes are on average replicated once per cell cycle, i.e. every few tens of minutes.

We consider in this hypothesis that C_α is a peaked function at 0, with a non-zero value beyond a small time τ such that it does not depend on wether a previous copy was the ancestor of the considered copy or not :

$$C_\alpha^{\mathrm{A}}(|t - t'|) = \langle \alpha^2 \rangle \times \tau \delta(t - t') + \langle \alpha_O \alpha_G \rangle (1 - \tau \delta(t - t')) - \langle \alpha \rangle^2,$$

This gives:

$$\langle P_a^2 \rangle_{\mathrm{A}} = \frac{9}{16} T^2 \langle \alpha_O \alpha_G \rangle (1 + 2\mathcal{R}_a + \mathcal{S}_{aa}) \langle (\overline{n_a})^2 \rangle + \frac{5}{16} \tau T (\langle \alpha^2 \rangle - \langle \alpha_O \alpha_G \rangle)(1 + 3\mathcal{R}_a) \langle \overline{n_a} \rangle. \tag{5}$$

In the hypothesis (B), we suppose that C_α is constant:

$$C_\alpha^{\mathrm{B}}(|t - t'|) = \langle \alpha^2 \rangle - \langle \alpha \rangle^2.$$

(We expect the actual form of C_α to be intermediate between those two, namely a smooth declining function on a time scale of a few minutes.) The hypothesis (B) is less realistic. It could correspond to mutations distinguishing different copies of a given gene. By noting that at any previous time each copy has exactly one ancestor, this translates in:

$$\sum_{i=1}^{n_a(t)} \sum_{i'=1}^{n_a(t')} \langle \alpha_a(i,t) \alpha_a(i',t') \rangle_{\mathrm{B}} = \langle \alpha_O \alpha_G \rangle n_a(t) n_a(t') + (\langle \alpha^2 \rangle - \langle \alpha_O \alpha_G \rangle)(n_a(t)\theta(t - t') + n_a(t')\theta(t' - t)),$$

where θ is the Heaviside function.

We then introduce a third quantity, $\mathcal{T}_a$, which is defined in the Appendixes, and can be shown to lay in the interval $[0, 9.9]$. (We will argue that this interval can be reduced to $[5.7, 6.1]$.) Then:

$$\langle P_a^2 \rangle_{\mathrm{B}} = \frac{9}{16} T^2 \langle \alpha_O \alpha_G \rangle (1 + 2\mathcal{R}_a + \mathcal{S}_{aa}) \langle (\overline{n_a})^2 \rangle + \frac{1}{8} T^2 (\langle \alpha^2 \rangle - \langle \alpha_O \alpha_G \rangle)(1 + \mathcal{T}_a) \langle \overline{n_a} \rangle. \tag{6}$$

The two hypotheses (A) and (B) thus only lead to different factors for the contribution of the average copy number [19]. This term is expected to be small, even in the hypothesis (B), where $1 + \mathcal{T}_a$ can be of the order of 10: the numerical pre-factor cancels it, one can expect $\langle \alpha^2 \rangle$ and $\langle \alpha_O \alpha_G \rangle$ to be of the same order of magnitude and, already for the plasmid of lowest copy number and for the chromosome, $\langle \overline{n_a} \rangle$ is significantly smaller than $\langle (\overline{n_a})^2 \rangle$. Moreover, if we let τ tend to the time of induction $2T$, we recover terms of the same order of magnitude, thus suggesting a low sensibility to the actual mathematical

translation of the hypotheses. The results will be presented and discussed with only the hypothesis (A), the more realistic, being considered; full computations with test functions confirmed that very close values for the PCN noise were found in the hypothesis (B) (data not shown).

IV. RESULTS AND DISCUSSION

By combining equations 3, 4 and 5 so as to eliminate the gene expression rates, and assuming that the replication of the chromosome is well controlled, we can now express the PCN noise:

$$\eta^2 = \left(\frac{1+\mathcal{R}_O}{1+\mathcal{R}_G}\right)\left(\frac{1+3\mathcal{R}_O}{1+3\mathcal{R}_G}\right)\left(\frac{1+2\mathcal{R}_G+\mathcal{S}_{GG}}{1+2\mathcal{R}_O+\mathcal{S}_{OO}}\right)\frac{\langle P_G\rangle}{\langle P_O\rangle} - 1 + \left(\frac{1+\mathcal{R}_O+\mathcal{R}_G+\mathcal{S}_{OG}}{1+2\mathcal{R}_O+\mathcal{S}_{OO}}\right)\frac{1}{\langle P_O P_G\rangle} \times \left\{\left(\frac{1+\mathcal{R}_O}{1+\mathcal{R}_G}\right)\frac{\langle P_G\rangle}{\langle P_O\rangle}\langle P_O^2\rangle - \left(\frac{1+3\mathcal{R}_O}{1+3\mathcal{R}_G}\right)\langle P_G^2\rangle\right\}. \quad (7)$$

Note that both the auto-correlation time τ introduced previously and the cell cycle length T have also been eliminated. Only terms related to replication and partition of genes, which we can bound, and the experimentally measured moments of protein numbers remain [20]. By making the conservative assumption that $\mathcal{R}_a$ and $\mathcal{S}_{ab}$ can independently take any value in their intervals, we can compute intervals in which the mean PCN per chromosome and the PCN noise are surely. They are presented in Table I for experiments at 37 °C, and in the Fig. 2 and Fig. 3 for various temperatures. We report both the intervals estimated with a general analysis and with a set of test functions for the moments of copy numbers. Values computed with the simple model are also shown. Both for means and noises, the values computed with the simple model fall in the middle of the intervals computed with the more realistic model.

As Fig. 2 shows, we can clearly distinguish the plasmids by their mean PCN per chromosome. Moreover, these results agree with previous, independent estimates, as discussed [4]. For the noises the picture is less clear. In the general study, the intervals found are too large for the results to be meaningful; but we know that we have largely overestimated them. In contrast, using test functions allows one to distinguish the plasmids by their PCN noises. In particular, we can notice that the partition system reduces the noise (compare mini-R1-*par*$^+$ and mini-R1-*par*$^-$), and that a plasmid with a high copy number (mini-ColE1) has a lower noise than a plasmid with a small copy number (mini-F), even though it has a partitioning system [21].

We tested the quality of the inference with simple computer simulations, where stochastic gene expression and plasmid replication were implemented (see the Appendixes for more details). Table II compares the true and inferred values of the mean PCN per chromosome

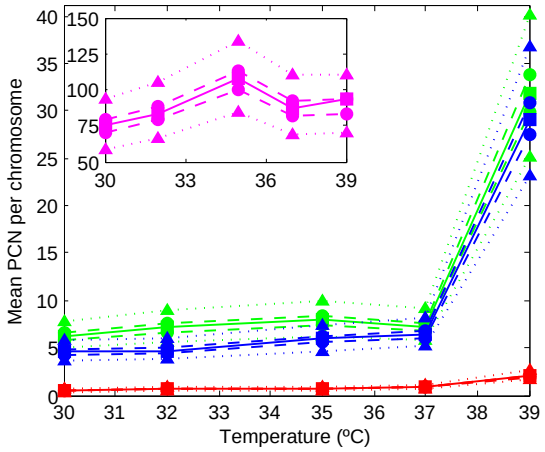


FIG. 2. (Color online) Mean PCN per chromosome $\langle \overline{n_O} \rangle / \langle \overline{n_G} \rangle$ at 30, 32, 35, 37 and 39 °C for the four plasmids studied here, from bottom to top: mini-F (red), mini-R1-*par*$^+$ (blue), mini-R1-*par*$^-$ (green), mini-ColE1 (magenta). The values obtained in three cases are plotted: with the simple model (squares, solid line), with the complete model and test functions (upper and lower bounds of the interval: circles, dashed lines) or within a general analysis (upper and lower bounds of the interval: triangles, dotted lines). The mini-R1 plasmids have a synthetic, thermo-sensitive origin of replication, the control of which is inactivated at high temperature.

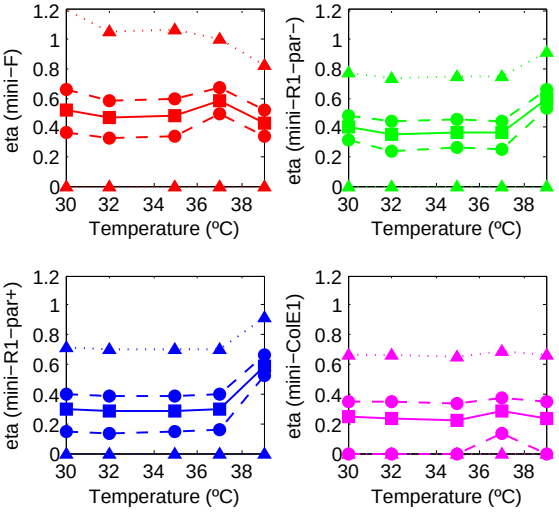


FIG. 3. (Color online) PCN noise η at various temperatures for the four plasmids studied here (see the caption of Fig. 2). We considered that cells divided once during the induction at 30 and 32 °C, twice at 35, 37 and 39 °C. Only the hypothesis (A) of short "memory" was considered. The results obtained with the simple model are fully recovered if we suppose a similar behavior for the plasmids and for the chromosome, see the main text.

TABLE I. Mean PCN per chromosome $\langle\overline{n_O}\rangle/\langle\overline{n_G}\rangle$ and PCN noise η computed with data from experiments at 37 °C, using the simple model or the complete one, either with a set of test functions or within a general analysis. Only the hypothesis (A) of short "memory" was considered. We assumed that cells divided twice during the induction.

	mini-F	mini-R1-*par*$^-$	mini-R1-*par*$^+$	mini-ColE1
$\langle n_O\rangle/\langle n_G\rangle$ simple	0.9	7.2	6.5	87
$\langle\overline{n_O}\rangle/\langle\overline{n_G}\rangle$ complete/test	[0.84, 0.95]	[6.8, 7.7]	[6.1, 6.9]	[82, 93]
$\langle\overline{n_O}\rangle/\langle\overline{n_G}\rangle$ complete/general	[0.71, 1.13]	[5.7, 9.1]	[5.1, 8.2]	[69, 110]
$\eta\times10^2$ simple	58	36	30	28
$\eta\times10^2$ complete/test	[50, 67]	[25, 45]	[16, 39]	[13, 38]
$\eta\times10^2$ complete/general	[0, 100]	[0, 74]	[0, 71]	[0, 68]

TABLE II. Test of the inference method with computer simulations, in four cases: 1. a synchronized population of bacteria with fixed division time, equal to half the induction time; 2. as 1. with an exponential age distribution; 3. as 2. with a distribution of division times, with mean equal to half the induction time; 4. as 3. with the mean division time equal to one third of the induction time.

	case 1	case 2	case 3	case 4
$\langle\overline{n_O}\rangle/\langle\overline{n_G}\rangle$ true	11.9	9.6	10.1	10.1
$\langle n_O\rangle/\langle n_G\rangle$ simple	12.8	10.1	11.0	11.7
$\langle\overline{n_O}\rangle/\langle\overline{n_G}\rangle$ complete/test	[12.0, 13.6]	[9.5, 10.7]	[10.4, 11.6]	[11.0, 12.4]
$\langle\overline{n_O}\rangle/\langle\overline{n_G}\rangle$ complete/general	[10.1, 16.1]	[8.0, 12.7]	[8.7, 13.8]	[9.3, 14.7]
$\eta\times10^2$ true	63	66	75	75
$\eta\times10^2$ simple	60	66	74	83
$\eta\times10^2$ complete/test	[54, 67]	[59, 72]	[68, 80]	[77, 89]
$\eta\times10^2$ complete/general	[0, 93]	[19, 98]	[35, 106]	[47, 114]

and the PCN noise in four cases, corresponding to different assumptions on the age and cell cycle duration distributions. In each case we find a very good agreement.

As it appears in equations 3, 4 and 5, what we call here "plasmid copy number", or "chromosome copy number", is precisely the average over one cell cycle of the number of copies of the gene coding for a fluorescent protein. A quantitative PCR (qPCR) measures $\langle n_O\rangle_{\rm q}/\langle n_G\rangle_{\rm q}$, where $\langle n_a\rangle_{\rm q} = \left\langle\int dt_0\, p(t_0)n_a(t_0)\right\rangle = \int dt_0\, p(t_0)\,\langle n_a\rangle(t_0)$. This quantity and the ratio $\langle\overline{n_O}\rangle/\langle\overline{n_G}\rangle$ reported here take in general different values. We have indeed noticed a discrepancy between the two approaches, but other explanations are likely [4].

We have made strong, but reasonable hypotheses on gene expression. We made intuitive notions explicit and gave them a well defined mathematical translation.

A deeper mathematical analysis could reduce significantly the general intervals found, but not below the intervals found with test functions. Here again, the experimental approach and derivation of the PCN mean and noise are self-consistent: there are no external inputs, even in the bounded "correction" quantities $\mathcal{R}_a$ or $\mathcal{S}_{ab}$, which depend only on the way the genes are replicated and inherited by daughter cells at divisions. Using the chromosome as a reference allowed us to get rid of global fluctuations: the number of divisions considered do not affect the results, fluctuations from proteins partition at division are suppressed, all fluctuations of gene expression are cancelled, and even the division time does not appear in the final result. This argues for the assumptions that the induction time is a multiple of the division time and that the variability in division times can be neglected not to affect the results. The simulations further confirm the robustness of this strategy, in particular the values inferred with the crudest assumptions ("simple model") are strikingly close to the true ones, both for the mean PCN per chromosome and the PCN noise (see Table II).

The only source of uncertainty that remains stems from the replication and partition of the plasmids and chromosome. The use of test functions suggests that it does not affect the results much. Moreover, if we suppose that both are similar, i.e. $\mathcal{R}_O \approx \mathcal{R}_G$ and $\mathcal{S}_{OO} \approx \mathcal{S}_{GG}$, we fully recover the simple model presented at the beginning and in the previous article [4].

The next obvious step would be to consider correlations *between* cells, which could in particular inform us on plasmids partition. Here however, we lack the information on the lineage (which cells share a common induced ancestor) necessary to make a practical use of these quantities.

The use of dual reporters to dissect sources of noise was first proposed and demonstrated in a simple framework: steady state of fully induced bacteria, with both reporters in as much a similar position as possible [22, 23]. Here we took a similar approach further, and made sense of an intuitive setup: by changing one element, namely the *locus* of insertion of the genes coding for fluorescent proteins, we were able to measure one particular source of noise. The analysis proposed here could serve as a model

for other derivations of this strategy.

Appendix A: Partition of proteins at cell divisions

Random partition of fluorescent proteins at cell divisions contributes only to the auto-correlation (fluctuations) of protein numbers. We suppose a binomial distribution of the number of inherited proteins. In the case of two divisions, this leads to adding the term $\frac{1}{12}\left(\frac{7-3\mathcal{R}_a}{1+\mathcal{R}_a}\right)\langle P_a\rangle$ to $\langle P_a^2\rangle$. In turn, this translates to adding the correction

$$\frac{\langle P_O\rangle}{12(1+\mathcal{R}_O)}\left(\left(\frac{1+3\mathcal{R}_O}{1+3\mathcal{R}_G}\right)(7-3\mathcal{R}_G)-(7-3\mathcal{R}_O)\right)$$

to $\langle P_O^2\rangle$, while leaving $\langle P_G^2\rangle$ unchanged, in the expression of the PCN noise η in the hypothesis (A) [14]. This term varies from $-0.2\langle P_O\rangle$ to $0.3\langle P_O\rangle$ when we independently vary $\mathcal{R}_O$ and $\mathcal{R}_G$ in the interval $[0.15, 0.45]$. Thus, with an expected number of proteins above 10 (probably hundreds to thousands), this correction is very small compared to $\langle P_O^2\rangle$ and can be safely neglected.

Appendix B: Estimation of $\mathcal{R}_a$, $\mathcal{S}_{ab}$ and $\mathcal{T}_a$

We briefly outline here the steps allowing us to bound $\mathcal{R}_a$, $\mathcal{S}_{ab}$ and $\mathcal{T}_a$. Full derivations can be found in [14].

We define:

$$\mathcal{R}_a = \frac{1}{\langle\overline{n_a}\rangle}\frac{1}{T}\int_0^T dt_0\, p(t_0)\int_0^{t_0} dt\,\langle n_a(t)\rangle.$$

We linearize the age distribution: $p(t_0) = (2\ln 2/T).2^{-t_0/T} \approx (2\ln 2/T)(1-\ln 2.t_0/T)$. There exists $t_0^* \in [T/2, T]$ such that

$$\mathcal{R}_a \approx \frac{1}{\langle\overline{n_a}\rangle}\frac{2\ln 2}{T^2}\left(1-\ln 2\frac{t_0^*}{T}\right)\int_0^T dt_0\int_0^{t_0} dt\,\langle n_a(t)\rangle.$$

We can suppose that $\langle n_a\rangle$ is increasing on $[0, T[$. It follows that $\int_0^T dt_0\int_0^{t_0} dt\,\langle n_a(t)\rangle \leq 1/2$. Recalling that $\overline{\langle n_a\rangle} = \langle\overline{n_a}\rangle$, this implies also: $\int_0^T dt_0\int_0^{t_0} dt\,\langle n_a(t)\rangle \geq \langle n_a(0)\rangle/(2\langle\overline{n_a}\rangle)$ and $\geq \langle\overline{n_a}\rangle/(2\langle n_a(T)\rangle)$. At steady state $\langle n_a(T)\rangle = 2\langle n_a(0)\rangle$. Thus, from the preceding inequalities:

$$\mathcal{R}_a \in [0.15, 0.45].$$

In the same way, linearizing $p(t_0)$ and showing that $\mathcal{S}_{ab}$ can be expressed in terms of the integral of a convex function, we find $\mathcal{S}_{ab} \in [0, 0.45]$.

In the case of two divisions, $\mathcal{T}_a$ is defined following:

$$\mathcal{T}_a = \frac{1}{\langle\overline{n_a}\rangle}\frac{1}{2T^2}\int_0^T dt_0 p(t_0)\left(7\int_0^T dt\,t + 27\int_0^{t_0} dt\,t\right.$$
$$\left.-3t_0\int_0^T dt + 16T\int_0^{t_0} dt - 3t_0\int_0^{t_0} dt\right)\langle n_a(t)\rangle.$$

We follow the same steps as before, only here we consider that each term can vary independently, thus highly overestimating the bounds for $\mathcal{T}_a$. We find $\mathcal{T}_a \in [0, 9.9]$.

Appendix C: Test functions

To gauge the quality of the previous estimates and fix minimal intervals, we computed $\mathcal{R}_a$, $\mathcal{S}_{ab}$ and $\mathcal{T}_a$ after postulating different shapes for the functions $\langle n_a\rangle$ and $\langle n_a n_b\rangle$.

The changes of variables $t \to t/T$ and $t_0 \to t_0/T$, and the normalization $\langle n_a\rangle \to \langle n_a\rangle/\langle n_a(0)\rangle$ leave $\mathcal{R}_a$ and $\mathcal{T}_a$ unchanged. We can thus limit ourselves to increasing functions on $[0, 1]$, going from 1 to 2. We considered step, sigmoid, exponential, logarithmic, sinus functions and monomials of various degrees. Each type is defined by one or two parameters: each parameter was given six values in the first case and four in the second case.

For $\mathcal{S}_{ab}$, we considered the product of any two functions among those above. (This implies $\langle n_a(T)n_b(T)\rangle = 4\langle n_a(0)n_b(0)\rangle$, which in general is not true.)

We used the exact expression of $p(t_0)$. We found:

$$\mathcal{R}_a^{\text{test}} \in [0.36, 0.44]; \mathcal{S}_{ab}^{\text{test}} \in [0.20, 0.28]; \mathcal{T}_a^{\text{test}} \in [5.7, 6.1].$$

Thus the interval found in general for $\mathcal{R}_a$ is rather good, whereas these results seem to confirm that the intervals found for $\mathcal{S}_{ab}$ and $\mathcal{T}_a$ were highly overestimated in the previous analysis.

Appendix D: Simulations

To test the method proposed in this article, we simulated roughly the experiment and compared the inferred quantities to the true ones, see Table II. The transcription of each reporter gene, the translation of the corresponding RNAs and their degradation, and the replication of the plasmids are each implemented as single stochastic reactions [24]:

$$\begin{aligned}\text{Gene}_a + \text{Inducer} &\longrightarrow \text{Gene}_a + \text{Inducer} + \text{mRNA}_a\\ \text{mRNA}_a &\longrightarrow \text{mRNA}_a + \text{Protein}_a\\ \text{mRNA}_a &\longrightarrow \emptyset\\ \text{Gene}_a &\longrightarrow 2\text{Gene}_a\end{aligned}$$

where a again refers to either the plasmid or the chromosome. We impose the number of chromosomes to go from 1 to 2 at 40% of the cell cycle. The number of inducers is also imposed: it is 0 until some time t_{lag}, and 1 until the end of the simulation.

The duration of the cell cycle T is either fixed (cases 1 and 2) or drawn from a gamma distribution with parameters chosen so that the average division time is 1800 or 1200 s, and typical variations of 10% (cases 3 and 4) [25]. The age of the cell at the beginning of the simulation is either 0 (case 1) or drawn from the exponential distribution indicated in the main text. The time t_{lag} has been

arbitrarily fixed at 10 times the mean cell cycle duration. The induction, i.e. the remainder of the simulation, has always been taken to last 3600 s.

We follow, in each simulation, one lineage. Upon cell division, the RNAs, proteins and plasmids are randomly picked to stay (their numbers drawn from a binomial distribution). 100,000 simulations were performed in each of the four cases.

The reaction constants were taken so that the obtained protein distributions (expressed from the chromosome) match those described by Swain et al. [22], with the same rates of transcription and translation for both reporters. Importantly, there is no control of the plasmid copy number: we simply fix the rate of replication to the growth rate $\ln 2/T$, so that the plasmids are on average replicated once per cell cycle. A steady state of plasmid copy number is thus not reached, contrary to the more realistic assumption made earlier; the fact that we can still recover the mean PCN and the PCN noise shows that even this assumption is not critical.

Each cell has 10 plasmids at the beginning of the simulation. The cell cycle average $\overline{\bullet}$ is taken during the first full cycle of induction.

All codes are available upon request.

[1] G. del Solar, R. Giraldo, M. J. Ruiz-Echevarria, M. Espinosa, and R. Diaz-Orejas, Microbiol. Mol. Biol. Rev., **62**, 434 (1998).

[2] K. Nordström and S. Dasgupta, EMBO Rep., **7**, 484 (2006).

[3] J. Paulsson and M. Ehrenberg, Quarterly Reviews of Biophysics, **34**, 1 (2001).

[4] J. Wong Ng, D. Chatenay, J. Robert, and M. G. Poirier, Phys. Rev. E, **81**, 011909 (2010).

[5] R. Y. Tsien, Annu. Rev. Biochem., **67**, 509 (1998).

[6] H. A. de Boer, L. J. Comstock, and M. Vasser, Proc. Nat. Acad. Sci. USA, **80**, 21 (1983).

[7] N. C. Shaner, R. E. Campbell, P. A. Steinbach, B. N. G. Giepmans, A. E. Palmer, and R. Y. Tsien, Nat. Biotechnol., **22**, 1567 (2004).

[8] This hypothesis is not necessary, supposing that it does not depend on time *on average* would lead to the same result, but it makes the notations simpler.

[9] J. Wong Ng, *Variation du Nombre de Copies de Plasmides au Sein de Populations Monoclonales de Bactéries*, Ph.D. thesis, École Normale Supérieure - Université Pierre et Marie Curie (Paris 6), Paris (2008).

[10] K. Skarstad, E. Boye, and H. B. Steen, EMBO J., **5**, 1711 (1986).

[11] J. A. Megerle, G. Fritz, U. Gerland, K. Jung, and J. O. Rädler, Biophys. J., **95**, 2103 (2008).

[12] P. Wang, L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun, Curr. Biol., **20**, 1099 (2010).

[13] F. C. Neidhart, ed., *Escherichia coli and Salmonella, Cellular and Molecular Biology*, 2nd ed., Vol. 2 (ASM Press, 1996) p. 1647.

[14] S. Ghozzi, *Dynamique d'expression d'un réseau de régulation génétique : la décision lyse / lysogénie chez le bactériophage Lambda*, Ph.D. thesis, École Normale Supérieure - Université Pierre et Marie Curie (Paris 6), Paris (2009).

[15] Here both α_a and n_a are particular realizations of rate and copy number, and thus are "noisy" functions. One can fix an arbitrarily small time step and consider the *initiations of transcription* to give a precise, well defined meaning to α_a without disregarding the discrete replication, transcription, translation and maturation steps.

[16] J. Elf, G.-W. Li, and X. S. Xie, Science, **316**, 1191 (2007).

[17] The fact that the promoter is present in many copies even argues for these dynamics to be faster.

[18] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Cell, **123**, 1025 (2005).

[19] The ratio of these factors is $\frac{5}{2}\left(\frac{1+3\mathcal{R}_a}{1+\mathcal{T}_a}\right)\frac{\tau}{T} \in [0.01, 0.2]$ if we consider that $\mathcal{R}_a$ and $\mathcal{T}_a$ are independent. (This interval is reduced to $[0.02, 0.03]$ if we let these two quantities vary in the intervals found with a set of test functions, see the Appendixes.).

[20] The number of proteins can be determined up to a constant multiplicative factor, the same for both colors (namely, the fluorescent intensity per EGFP molecule in the selected green channel) [4]; here, the contribution of random partition at divisions having been neglected, only ratios of the same order show up, thus canceling this unknown factor.

[21] We notice however that for the mini-R1 plasmids, both averages and noises increase at high temperature; this could come from fluctuations in the number of mature thermo-sensitive replication control proteins.

[22] P. S. Swain, M. B. Elowitz, and E. D. Siggia, Proc. Nat. Acad. Sci. USA, **99**, 12795 (2002).

[23] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science, **297**, 1183 (2002).

[24] D. F. Anderson, J. Chem. Phys., **127**, 214107 (2007).

[25] J. L. Lebowitz and S. I. Rubinow, J. Math. Biol., **1**, 17 (1974).